DISCOVERING SENTENCES FOR ARGUMENTATION ABOUT THE MEANING OF STATUTORY TERMS

by

Jaromir Savelka

JUDr. in Law and Legal Science, Masaryk University, 2013

M.S. in Intelligent Systems, University of Pittsburgh, 2016

Submitted to the Graduate Faculty of

the School of Computing and Information in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2020

UNIVERSITY OF PITTSBURGH

SCHOOL OF COMPUTING AND INFORMATION

This dissertation was presented

by

Jaromir Savelka

It was defended on

April 20, 2020

and approved by

Kevin D. Ashley, School of Law, LRDC

Milos Hauskrecht, School of Computing and Information

Daquing He, School of Computing and Information

Diane J. Litman, School of Computing and Information, LRDC

Dissertation Director: Kevin D. Ashley, School of Law, LRDC

DISCOVERING SENTENCES FOR ARGUMENTATION ABOUT THE MEANING OF STATUTORY TERMS

Jaromir Savelka, PhD

University of Pittsburgh, 2020

In this work I studied, designed, and evaluated computational methods to support interpretation of statutory terms. Understanding statutes is difficult because the abstract rules they express must account for diverse situations, even those not yet encountered. The interpretation involves an investigation of how a particular term has been referred to, explained, interpreted, or applied in the past. This is an important step that enables a lawyer to then construct arguments in support of or against particular interpretations. Going through the list of results manually is labor intensive. A response to a search query may consist of hundreds or thousands of documents. I investigated the feasibility of developing a system that would respond to a query with a list of sentences that mention the term in a way that is useful for understanding and elaborating its meaning. I treat the discovery of sentences for argumentation about the meaning of statutory terms as a special case of ad hoc document retrieval. The specifics include retrieval of short texts (sentences), specialized document types (legal case texts), and, above all, the unique definition of document relevance.

This work makes a number of contributions to the areas of legal information retrieval and legal text analytics. First, a novel task of discovering sentences for argumentation about the meaning of statutory terms is proposed. The task includes analyzing past treatment of a statutory term, a task lawyers routinely perform using a combination of manual and computational approaches. Second, a data set comprising 42 queries (26,959 sentences) was assembled to support the experiments presented here. Third, by systematically assessing the performance of a considerable number of traditional information retrieval techniques,

I position this novel task in the context of a large body of work on ad hoc document retrieval. Fourth, I assembled a unique list of 129 descriptive features that model the retrieved sentences, their relationships to the terms of interest, as well as the statutory provisions they come from. I demonstrate how the proposed feature set could be utilized in learning-to-rank settings by showing how a number of machine learning algorithms learn to rank the sentences with very reasonable effectiveness. Fifth, I analyze the effectiveness of fine-tuning pre-trained language models in the context of this special task and demonstrate a very promising direction for future work.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

Above all, I would like to thank to my advisor Kevin D. Ashley who has been remarkably supportive with regards to all aspects of my PhD life. There is no other person in the world who had more influence on my professional development than he. It was a great pleasure to be Kevin's student and I consider myself extremely lucky. I would like to thank to the members of my dissertation committee, Milos Hauskrecht, Daqing He, and Diane J. Litman, who have volunteered their time to help me with the final steps of my graduate studies. I would also like to thank Rebecca Hwa who supported me by sitting on my preliminary and comprehensive exam committees.

The Intelligent Systems Program (ISP) is a fantastic environment for anyone interested in applications of AI to specific domains such as Law. I would like to thank the program administrators, Wendy Bergstein and Michele L. Thomas, who always provided great advice and support. I am also very thankful for the great courses I was able to take as an ISP student. I had little to no prior experience with applications of ML and NLP. And while all the courses contributed to my development I would especially like to mention Janyce Wiebe's Intro to AI, Diane Litman's Intro to NLP, and Milos Hauskrecht's two ML courses, that I believe were instrumental in my transition.

During my PhD studies I had the pleasure to work with a number of inspiring collaborators. Among many I would specifically like to mention Hannes Westermann who has made my work during the last year so much more enjoyable. Similarly, a huge thanks goes to Matthias Grabmair whom I probably owe more than I realize. On a variety of projects, I especially enjoyed working with Vern R. Walker, Jakub Harašta, Prakash Poudyal, Gaurav Trivedi, Mohammad Falakmasir, and Huihui Xu.

There were a number of people who played an important role in my academic development

before coming to ISP. A lot of what I learned from them continues to have significant impact on my work. Although, it is impossible to mention everyone I would definitely like to thank to Radim Polčák, Martin Škop, and Zdeněk Říha for their mentorship. Furthermore, I really enjoyed working with Michał Araszkiewicz, Matěj Myška, Michal Koščík, and Terezie Smejkalová.

Apart from the ISP, several other institutions and companies supported my research. The Centers for Disease Control and Prevention sponsored my early work through the University Research Council Multidisciplinary Small Grant Program. The National Institute of Justice provided me with 2 years of funding directly dedicated to the work on this dissertation through their Graduate Research Fellowship in Science, Technology, Engineering, and Mathematics. University of Pittsburgh's Institute for Cyber Law, Policy, and Security funded the data gathering activity with the award of one of the Pitt Cyber Accelerator Grants. Reed Smith, LLP and Gravity Stack, LLC have my gratitude for being supportive and understanding of my ambition to engage in graduate studies.

Finally, but most importantly I would like to thank to my family. Being a PhD student in a foreign country means that my wife, my children, and I have been apart from many people we love the most. This is not only difficult for us but it places a burden on them as well. I would specifically like to thank our parents, Jaroslava Šavelková, Božena Počtová, and Miroslav Počta, and my brother, Petr Šavelka, for their patience, support, and understanding. Not being with them was the most difficult part of my studies. Most of all, I would like to thank to my wife Jana. During the whole time she has always stood by me. This means that she has placed my ambition to pursue the PhD studies over her own career desires and preferences. This means that she had to leave her family and closest friends to follow me to the United States. This means that she, with embarassingly little help from me, takes care of our two wonderful sons, Filip and Daniel. And she never complained (much). I know that this period of my life would have been much less fun if I had to go through it alone. For all this and much more I will always be grateful to my wife. And now I am looking forward to what comes next knowing that we will go through it together as always.

1.0 INTRODUCTION

In this work I study, design, and evaluate computational methods to support interpretation of statutory terms. In legal argumentation a lawyer must often defend a specific account of the meaning of one or more terms (i.e., words, phrases). The persuasiveness and validity of a complex argument may hinge on a particular account of the meaning. Argumentation about the meaning of a term may even be the crux of an overall argument. Understanding statutes is difficult because the abstract rules they express must account for diverse situations, even those not yet encountered. The legislators use vague, open textured terms, abstract standards, principles and values in order to achieve generality at the cost of uncertainty. Consider the following (abridged) excerpt from 29 U.S. Code § 203:

> "Enterprise" means the *related activities* performed [...] by any person or persons for a *common business purpose* [...]

A lawyer wishing to argue that two restaurants located in different parts of a city owned by a single person do not constitute an enterprise may, e.g., argue that they cannot be considered *related activities* or that their operation is not performed for a *common business purpose*. This effectively amounts to defending an account of *common business purpose* where the common ownership of the two restaurants could not be subsumed under it (similarly with respect to *related activities*).

The interpretation involves an investigation of how the term has been referred to, explained, interpreted or applied in the past. This is an important step that enables a lawyer to then construct arguments in support of or against particular interpretations. Searching through a database of statutory and regulatory texts, case law, legislative history, or law review articles one may stumble upon sentences such as these:

i. [. . .] the fact of common ownership of the two businesses clearly is not sufficient to establish a *common business purpose*.

ii. [. . .] the profit motive is a *common business purpose* if shared.

iii. Were the buildings managed by their owners, the Government would not attempt to link them together as an enterprise bound together by a '*common business purpose*.'

iv. Because the activities of the two businesses are not related and there is no *common business purpose*, the question of common control is not determinative.

v. The defendants weakly challenge the *common business purpose* conclusion [. . .]

Some of the sentences are useful for the interpretation of the term *common business purpose* from the example provision (i. and ii.). Some of them look like they may be useful (iii.) but the rest appear to have very little if any value (iv. and v.). Going through the sentences manually is labor intensive. A response to a search query may consist of hundreds or thousands of documents. Usually most of the sentences that contain the term of interest would be useless and redundancy would be high.

In this work I investigate the feasibility of developing a system such as the one shown in Figure 1. Using such a system, a user reading a statutory provision could indicate his or her interest in a specific word or a phrase the provision contains. The system would then respond with a list of short text snippets, such as sentences, that mention the phrase in a way that is useful for understanding and elaboration on its meaning. In case of the *common business purpose* a lawyer could use sentence i. to argue that the two restaurants cannot be considered an enterprise:

> Since *the common ownership of the two businesses is not sufficient to establish a common business purpose* (sentence i.) it is not possible to conclude that the two restaurants share the common business purpose. Therefore, the two restaurants cannot be considered an enterprise.

An opposing lawyer could use sentence ii. to argue the opposite:

> The two restaurants share the profit. Since *the profit motive is a common business purpose if shared* (sentence ii.) it is possible to conclude that the two restaurants share the common business purpose. Therefore, the two restaurants may be considered an enterprise.

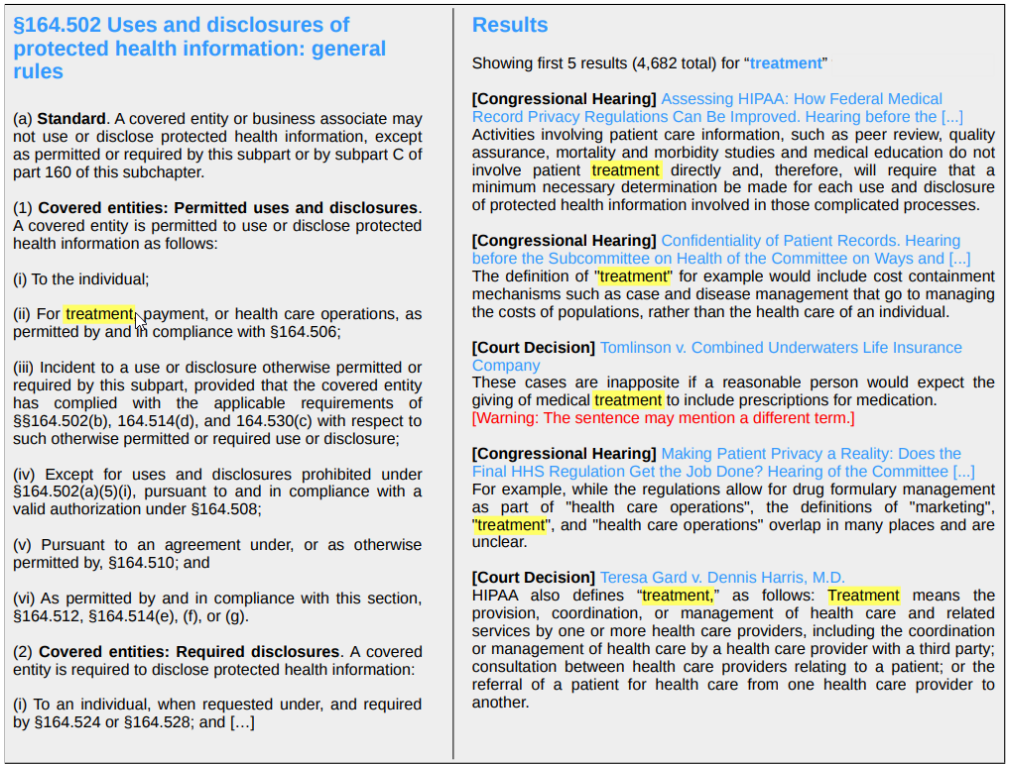

Figure 1: The Figure shows a mock-up interface with an example statutory provision on the left (§ 164.502 HIPAA). The user indicated that he or she is interested in the meaning of the term "treatment" as used in the provision (highlighted in yellow). The system responds with a list of sentences that are useful for the interpretation of the term. The system may even warn the user if it determines that a sentence may mention the term in a slightly different meaning.

I treat the discovery of sentences for argumentation about the meaning of statutory terms as a special case of ad hoc document retrieval which is one of the traditional tasks in the field of Information Retrieval (IR). The specifics include retrieval of short texts (i.e., sentences as opposed to, e.g., full web pages in general web search), specialized document types (i.e., statutory provisions and legal case texts), and above all a unique definition of document relevance (i.e., the utility of a sentence for the argumentation about the meaning of a term of interest). My work could then be understood as a systematic analysis of this special case

of ad hoc document retrieval resulting in a proof of concept system facilitating the discovery of relevant sentences.

The work proceeds via systematic assessment of thesis statements laid down in the following section (1.1). The overall goal is to provide empirical evidence that a computer system is capable of identifying the useful sentences automatically. At first, I investigate how well one could model sentences' relevance using a number of well-established strategies based on measuring similarity between the sentence and the term of interest. In order to address some of the shortcomings, the effects of taking sentences' contexts into account are analyzed. This leads to a measure which captures the definition slightly better. Furthermore, several other approaches modeling different aspects of the relevance definition are examined. These are the amount of additional information a sentence provides over the source provision (novelty), as well as the topical similarity between the source provision and the full text of a case a sentence comes from. Finally, I show that a compound measure based on all the above appears to be a reasonable model of the sentences' usefulness.

The second part of this work is focused on tackling the task in classical learning-to-rank settings. Using the insights gained from the previous stage of the work I have compiled a list of 129 descriptive features. A number of traditional ML algorithms are trained and evaluated on the data set consisting of the sentences represented in terms of these features. This shows that a sentence classification or regression system for predicting the value of a sentence could be trained as a reasonable model for ranking the sentences according to their utility. I also experimented with casting the task as ordinal classification and pair-wise relevance classification. Overall, this approach appears to model the relevance definition much better than the hand-crafted compound measure.

The final part is focused on the use of pre-trained language models based on deep neural network architectures. I show that fine-tuning of such models for the special task of retrieving sentences for argumentation about the meaning of statutory terms appears to be a promising approach. This leads to a learning-to-rank system that does not require hand-crafted features. Three variants of the model are considered. The simplest one considers the sentences themselves and predicts the usefulness based solely on their texts. The second variant models the relationship between the term of interest and a sentence. The third model

focuses on the relationship between the source provision and a sentence. The two models for classification of the textual pairs turn out to be very promising.

Finally, it is important to mention that I have made several simplifying assumptions that make the work feasible. The first such assumption concerns the set of relevant sentences. Specifically, only those sentences that have at least one exact mention of the term of interest are considered potentially relevant in this work. Secondly, due to limited resources it was not possible to include every potential term of interest into the data set. I limited the terms to those that returned fewer than 5,000 results (sentences). Furthermore, I assumed that the best length for a passage elaborating on the meaning of a statutory term is a sentence. This is not always true. Sometimes the best passage could be just a part of a sentence. Other times it could be multiple sentences or even a whole paragraph. In this work I retrieve sentences from the opinions of the courts only. However, other types of documents contain useful sentences as well. Legislative histories and legal commentaries tentatively appear to be very promising.

1.1 THESIS STATEMENTS

The aim of this work is to support the following general statement:

Given a statutory provision, a phrase in that provision, and a database of case law, a computer system can autonomously rank the sentences retrieved from the case law in terms of how useful they are for argumentation about the meaning of the phrase.

This goal is being achieved by assessment of several sub-statements described below.

1.1.1 Features Indicative of Sentences' Usefulness

A number of features indicative of sentences' usefulness may be identified. This is tested with the hypotheses listed below under S1.–S5.. The features describe different aspects of the computational definition of usefulness I implement in this work. This means that several

features provide even stronger indication when used in combination. This is tested with the hypotheses under S6.

S1. A similarity between the phrase and a retrieved sentence is indicative of sentence's relative usefulness for argumentation about the meaning of the phrase.

S1.1 This holds for similarity measures that rely on matching only the phrase terms while ignoring other terms in the sentence (e.g., BM25).

Null: Rankings produced by such methods do not differ from random rankings.

S1.2 This holds for similarity measures based on word embeddings that consider the relationship between the phrase terms and all the terms in the sentence.

Null: Rankings produced by such methods do not differ from random rankings.

S1.3 Using a linear combination of a method from S1.1 and a method from S1.2 one can obtain even stronger indication of sentences' utility.

Null: Rankings produced by combined methods do not differ from rankings produced by their individual components.

S2. A similarity between the phrase and a retrieved sentence, *including its context*, is even more indicative of sentence's relative usefulness for argumentation about the meaning of the phrase than just a similarity to the sentence itself.

S2.1 This holds for similarity measures that rely on matching only the phrase terms while ignoring other terms in the sentence and its context.

Null: Rankings produced by such methods applied to a sentence including its context do not differ from rankings produced by applying them to the sentence only.

S2.2 This holds for similarity measures based on word embeddings that consider the relationship between the phrase terms and all the terms in the sentence and its context.

Null: Rankings produced by such methods applied to a sentence including its context do not differ from rankings produced by applying them to the sentence only.

S2.3 Using a linear combination of a method from S2.1 and a method from S2.2 one can obtain even stronger indication of sentences' utility.

Null: Rankings produced by combined methods applied to a sentence including

its context do not differ from rankings produced by applying their individual constituents.

S3. The novelty of a sentence with respect to the statutory provision the phrase comes from (i.e., the additional information the sentence provides over what is already known from the provision) is indicative of sentence's relative usefulness for argumentation about the meaning of the phrase.

S3.1 Lack of sentence's novelty indicates low utility.

Null: Rankings produced by placing sentences with low novelty scores to the end of the results list do not differ from random rankings.

S3.2 High novelty is indicative of high utility.

Null: Rankings produced by ordering the full results list from the most novel sentences to the least do not differ from rankings produced by the methods from S3.1.

S4. A topical similarity of a full case text, from which the retrieved sentence comes, to the source provision is indicative of the sentence's relative usefulness for argumentation about the meaning of the phrase.

S4.1 Lack of the case text's topical similarity indicates low utility.

Null: Rankings produced by placing sentences that come from cases with low topical similarity to the end of the results list do not differ from random rankings.

S4.2 High topical similarity is indicative of high utility.

Null: Rankings produced by ordering the full results list from the sentences coming from the most topically similar cases to those that come from the least similar cases do not differ from rankings produced by the methods from S4.1.

S5. The functional part of a case text where a sentence appears is indicative of the sentence's relative usefulness for argumentation about the meaning of the phrase.

Null: Rankings produced by placing sentences that come from a specific functional part at the beginning of the results list do not differ from random rankings.

S6. Methods from S1.–S5. can be integrated in such a way that their aggregate output is more indicative of sentence's relative usefulness for argumentation about the meaning of the phrase than the output of the individual constituents.

S6.1 This holds for integrated methods where the core similarity measure relies on matching the phrase terms only while ignoring other terms in the sentence and its context augmented with the methods from S3.1 and S4.1.

Null: Ranking produced by such integrated methods do not differ from rankings produced by their individual constituents.

S6.2 This holds for integrated methods where the core similarity measure relies on word embeddings that consider the relationship between the phrase terms and all the terms in the sentence and its context augmented with the methods from S3.1 and S4.1.

Null: Rankings produced by such integrated methods do not differ from rankings produced by their individual constituents.

1.1.2 Feature Engineering-based Learning-to-Rank Approaches for Retrieval of Useful Sentences

Given the set of features indicative of sentences' usefulness, the task of retrieving sentences can be successfully tackled as a learning-to-rank problem. This means that the features enable autonomous learning of a ranking function, which is a reasonable model of sentences' utility for argumentation about the meaning of statutory terms. This is tested with the hypotheses under S7.

S7. Using the set of features put together on the basis of investigating S1.–S5., a sentence classification or regression system for predicting the value of a sentence could be trained as a reasonable model for ranking the sentences according to their utility.

S7.1 Predictions of some standard ML classification or regression models trained on the said features to predict sentences' value can be utilized as a basis for sophisticated ranking that takes into account a variety of different features.

Null: Rankings produced by such methods do not differ from the rankings produced by the methods associated with thesis statement S1.1 (using BM25) and with thesis statement S2.1 (using BM25-c).

S7.2 Encoding the labels in such a way that the system treats them as ordinal leads to an improvement in performance when compared to S7.1.

Null: The performance of the methods trained on the ordinal labels does not differ from the performance of the methods from S7.1.

S7.3 The ranking model can be improved by transforming the classification problem into a pair-wise ranking task.

Null: The performance of the methods trained in the pair-wise ranking settings does not differ from the performance of the methods from S7.1

1.1.3 Fine-tuning Pre-trained Language Models for Retrieval of Useful Sentences

Given the proven ability of pre-trained language models based on deep neural network architectures to learn higher-level features in the process of their fine-tuning for a specific task, the sentence retrieval can be successfully tackled as a learning-to-rank problem with no need for hand-crafted features. This is tested with the hypotheses under S8.

S8. Using the original textual inputs (i.e., the phrase, the provision it comes from, and the sentences) a sentence classification system for predicting the value of a sentence in multi-class settings could be trained as a reasonable model for ranking the sentences according to their utility.

S8.1 Predictions of some pre-trained language model, fine-tuned on the task of sentence classification in terms of their usefulness for argumentation about the meaning of a phrase from statutory provision, can be utilized as a basis for sophisticated ranking.

Null: Rankings produced by such methods do not differ from the rankings produced by the methods associated with thesis statement S1.1 (using BM25) and with thesis statement S2.1 (using BM25-c).

S8.2 Predictions of some pre-trained language model fine-tuned on the task of sentence pair classification between a phrase from statutory provision and a sentence with respect to the usefulness can be utilized as a basis for sophisticated ranking.

Null: Rankings produced by such methods do not differ from the rankings produced

by the methods associated with thesis statement S1.1 (using BM25) and with thesis statement S2.1 (using BM25-c).

S8.3 Predictions of some pre-trained language model fine-tuned on the task of sentence pair classification between a statutory provision and a sentence with respect to the usefulness can be utilized as a basis for sophisticated ranking.

Null: Rankings produced by such methods do not differ from the rankings produced by the methods associated with thesis statement S1.1 (using BM25) and with thesis statement S2.1 (using BM25-c).

1.2 CONTRIBUTIONS

This work makes a number of contributions to the areas of legal information retrieval and legal text analytics. Like other more recent contributions to these areas, this work belongs to the general stream of applied ML and NLP. The most notable contributions include:

- **Novel Task Definition**: This work proposes a novel task of discovering sentences for argumentation about the meaning of statutory terms. The task corresponds to the analysis of past treatment of a statutory term, which lawyers routinely perform using a combination of manual and computational approaches. The analysis is a key component in the overall activity of arguing a particular point that involves uncertainty about the meaning of a statutory phrase. The task is framed as a *sentence ranking problem*. The goal is to rank the sentences from the most to the least valuable ones. The value of a sentence is defined on an ordinal scale with four levels. Detailed annotation guidelines that capture the definition of sentences' value for argumentation about the meaning of statutory terms were developed.
- **Data Set**: To support the experiments presented in this work a statutory interpretation data set was assembled. The data set comprises 42 queries (i.e., statutory phrases) each of which is associated with a number of sentences retrieved from a sizable corpus of the United States case law. Each of the 26,959 sentences was seen by at least three human annotators (14 annotators in total). The data set will be released to the public.

- **Task Analysis**: I have systematically analogized the task to several well established IR problems. Assessing the effectiveness of a considerable number of traditional techniques to address those problems I show that the task is related to document relevance ranking, novelty detection, as well as topic modeling. By doing this I was able to position this novel task in the context of a large body of work on ad hoc document retrieval. A combination of several techniques is proposed as a specialized retrieval framework that significantly outperforms the random baseline.

- **Innovative Hand-crafted Feature List**: Based on the detailed task analysis, I assembled a list of 129 descriptive features that model the retrieved sentences, their relationships to the statutory phrase as well as the provision of law it comes from. These features are a domain specific adaptation of similar lists from areas such as general web search. I showed that the proposed feature set could be successfully utilized in learning-to-rank settings by demonstrating how a number of algorithms learn to rank the sentences with very reasonable effectiveness.

- **Effectiveness of Deep Representations**: This work can be understood as a detailed case study on a task where shallow document representations fail to model the problem adequately. At the same time, a deep representation appears to be capable of capturing the useful signal potentially eliminating the need for hand-crafted features. The whole work then presents a compelling demonstration of the effectiveness of methods based on deep representations in a specific task from a specific domain. This is important because advances in general NLP and ML do not always fully transfer to specialized domains such as legal texts.

Tentatively, it appears the work may also provide some valuable contributions to areas other than legal IR and legal text analytics. For example, the work could be understood as a special case of argument component identification and as such it could be very interesting to the field of argument mining. Similarly, as the work involves automation of investigating meaning of concepts it appears there might be certain aspects that would be of great interest to the area of education and learning science. However, I leave these streams of potential contributions largely unexplored at the moment and plan to revisit them in future work.

1.3 OUTLINE

Chapter 2 provides a very gentle introduction to statutory law. It explains how language's imprecision manifests when general rules encoded in statutes need to be applied to factual circumstances. Specifically, this imprecision necessitates the interpretation or argumentation about the meaning of certain terms embedded in statutory provisions. The chapter then elaborates on the role the analysis of the term's past treatment plays in the argumentation and how it is being performed. In Chapter 3 the task of retrieving case law sentences that are useful for argumentation about the meaning of statutory terms is described. The task is placed in the larger context of general legal problem solving, and application of law to factual circumstances specifically. The individual elements such as the term of interest, the source provision, or the case law database are introduced.

In Chapter 4 the effort of putting together the data set is outlined. The chapter presents details on how the annotation was performed beginning with the draft of annotation guidelines and training of the annotators and concluding with the adjudication of disagreements. The outcome of the process is analyzed in detail and some apparent limitations are exposed and discussed. Finally, the resulting data set is described. Chapter 5 then lays down the blueprint for the experiments performed in the following chapters.

Chapter 6 presents the analysis of applying different methods traditionally used in ad hoc document retrieval to the task of retrieving useful sentences. The numerous experiments with methods assessing similarity between the document and the query, novelty, topical relatedness, as well as the effects of taking the context into account, serve three main purposes. First, it is established that these methods are able to model different aspects of the task, affirming that it is viable to understand it as a special case of ad hoc document retrieval. Second, the detailed understanding of the task enabled me to propose a simple compound system that performs the sentence retrieval task with an effectiveness which appears to be higher than that of its individual constituents. Finally, I used the insights gained from the analysis to put together a list of features that could be used in the learning-to-rank setup.

In Chapter 7 the features designed on the basis of the experiments presented in Chapter 6 are described. The experiments with different setups of using a number of off-the-shelf

ML algorithms to learn the ranking function from the features are presented. These include a point-wise approach where the task is modeled as multi-label document classification as well as regression. A special transformation of the task into ordinal classification is applied. Finally, the pair-wise approach where the task is re-cast into binary document pair classification is considered as well.

Chapter 8 presents my initial attempts to tackle the task by fine-tuning pre-trained language models based on deep neural network architectures. Three different setups are assessed and compared to analogical conditions using different subsets of hand-crafted features described in Chapter 7. The initial results are very promising and suggest a path forward that does not require hand-crafted features and could potentially lead to superior performance.

Chapter 9 provides a discussion that spans over what was covered in Chapters 6–8. Here, an overall trend of improving performance as more sophisticated methods are used becomes apparent. First, the selection of methods that are discussed in this chapter is explained. Second, the chapter elaborates on the progress in performance as the focus moves from simpler measures based on similarity or novelty and their combination to more sophisticated learning-to-rank methods based on the feature-engineering approach and pre-trained language models.

2.0 BACKGROUND

2.1 STATUTORY LAW

Statutes are written laws enacted by legislative bodies. They set forth the body of legal norms which are legally binding rules of conduct. The compilation of statutes is typically organized into a hierarchical structure. For example, the United States Code (U.S.C.) is the official collection of the federal statutes of the United States. It comprises 54 titles; each of them subdivided into a number of chapters. Intermediary levels such as parts, subtitles, or divisions may be used. A single chapter is concerned with a specific domain of legal regulation and comprises sections that contain provisions expressing the individual legal rules (e.g., rights, prohibitions, duties). A simple rule on delivery of mail to persons that are not residents of the place of address is expressed like this:

> Whenever the Postal Service determines that letters or parcels sent in the mail are addressed to places not the residence or regular business address of the person for whom they are intended, to enable the person to escape identification, the Postal Service may deliver the mail only upon identification of the person so addressed.
>
> (39 U.S.C. § 3004)

A placement of the rule in the hierarchy of U.S.C. is shown in the top part of Figure 2. Statutory law does not have to comprise rules exclusively. Definitions such as the following are commonly used (the middle part of Figure 2):

> The term "assets" includes contracts, facilities, property, records, unobligated or unexpended balances of appropriations, and other funds or resources (other than personnel).
>
> (6 U.S.C. § 101)

In preambles of statutes or in other places it is possible to encounter declarations of goals or

```
Title 39.  Postal Services
└─Part IV. Mail Matter
  └─Chapter 30.  Nonmailable Matter
    └─Section 3004.  Delivery of mail to persons not residents of the
      place of address
      Whenever the Postal Service determines that letters [...]


Title 6.  Domestic Security
└─Chapter 1.  Homeland Security Organization
  └─Section 101.  Definitions
    The term ‘‘assets’’ includes contracts, facilities, property, [...]


Title 49.  Transportation
└─Subtitle V. Rail Programs
  └─Part C. Passenger Transportation
    └─Chapter 241.  General
      └─Section 24101.  Findings, Mission, and Goals
        Rail passenger transportation can help alleviate [...]
```

Figure 2: The Figure shows three example sub-structures of the U.S. Code. The top part shows the placement of the rule on delivery of mail to persons that are not residents of the place of address. The middle part shows an example definition of "assets." The bottom part presents an example provision expressing a goal.

values (the bottom part of Figure 2):

> Rail passenger transportation can help alleviate overcrowding of airways and airports and on highways.
>
> (49 U.S.C. § 24101)

Provisions of law are difficult to understand because the rules they express must account for diverse situations, even those not yet encountered. This means the rules need to be abstract and general. In the words of Herbert L. A. Hart, provisions of law need to communicate general standards and refer to classes of persons, and to classes of acts, things, and circumstances. [44, p. 124] In order to achieve the required generality, legislators use vague [29] open textured [44] terms, abstract standards [28], principles, and values [20]. Hence, the provisions may appear imprecise and unclear. An understanding of a single provision typically depends, among other things, on well-developed knowledge of the meaning of its constituent terms and phrases. Doubts about the meaning of a provision may be removed

by interpretation. [71] The successful use of the rules depends on a capacity to recognize particular acts, things, and circumstances as instances of the general classifications which the law makes. [44, p. 124–126] In other words, in order to use the rules successfully it is necessary to map the general norms onto specific factual circumstances. This may often prove to be a considerable challenge.

2.2 IMPRECISION IN LANGUAGE

The difficulties in understanding statutory provisions are but a specific, and perhaps apparent, manifestation of language's imprecision. Natural language is inherently imprecise. This applies even to common terms such as "red," "bald," and "young." For the sake of example, let us consider vagueness as discussed in [27] as one of the possible causes of imprecision. Claiming that a term is vague usually amounts to ascribing it with three related features:

1. the existence of borderline cases,
2. the lack of a sharp boundary along the transition from clear cases to clear counter-instances, and
3. susceptibility to sorites arguments (see below).

Considering the example of the term "young." Almost everyone could agree that a 16-year-old man is "young" while a 90-year-old man is not. It is much less clear if a 30, 40, or 50-year-old man could be considered "young." Instances that are unclear with respect to their membership in a category are the borderline cases. Assuming a 16-year-old man is "young" while a 90-year-old man is not, it would be interesting to identify a boundary age separating instances that belong to the category from those that do not. Although it is clear that the boundary is an age greater than 16 and less than 90 ($16 < \text{boundary} < 90$), it turns out that it is not really possible to select a single age as the boundary. If almost everyone can agree that a man of a certain age is "young" then almost everyone should also agree that a man who is one year older is "young" as well. Although, this seems reasonable a repeated application of this logic could lead to an obviously absurd conclusions. For example,

starting with a 16-year-old man as "young" it is possible to conclude that a 90-year-old man is "young" too. This phenomenon is called a sorites paradox. [27, pp. 12]

Vagueness is not restricted to adjectives. It is found in many other lexical categories for which some notion of grading can be relevant. Nouns such as "heap," or even "chair" or "apple" can be vague. Verbs such as "run" or "walk" could be vague too. We can even consider determiners as vague, such as "many," "few," "much," or "little" as well as adverbs (e.g., "quickly," "surprisingly," "clearly") and modifiers (e.g., "very," "somewhat," "completely"). [27, p. 3]

There are numerous other properties of language, similar or related to vagueness, that result in language's imprecision. Some examples are underdetermination, openness of meaning, contextual variability, inexactness, overdetermination, overlap or ambivalence between categories, [27, pp. 7–8] ambiguity, and generality. [112] The fine-grained distinctions between the properties are not crucial for this work. The main point is that a language communication does not have a universal and precise meaning everyone could agree on. This does not appear to be a flaw in the communication but it rather appears to be its feature.

The imprecision does not seem to cause major problems in everyday communication. Let us consider the following example utterance:

> It is cold outside. Wear warm clothes.

In everyday communication this is a perfectly meaningful utterance. Any reasonable person could understand it and, moreover, a person could act upon it (e.g., by wearing a sweater). One could object that the utterance may in fact be precise given the context in which it is uttered (e.g., the specific weather conditions, the available clothes, the clothing habits of the person to whom the utterance is directed). This is certainly true and in many circumstances the utterance may be as precise as pointing to a specific piece of clothing. For example, there might be a mutual understanding between the persons involved in the discussion that the utterance requests a specific sweater to be worn. However, this merely means that the utterance conveys a message that can be acted upon. The utterance itself is still imprecise.

Let us consider the term "cold"—assuming it is 38°F outside we can easily show the term manifests the three features described above. While almost everyone would agree that 90°F

is not cold, 45°F or 50°F are examples of the boundary cases. The lack of a sharp boundary as well as the susceptibility to the sorites argument are quite clear too. The same applies to other terms such as "warm," "clothes," or even "outside." The main point is that a perfectly valid and useful everyday communication could be achieved with terms and phrases that are imprecise.

2.3 LANGUAGE'S IMPRECISION AND APPLICATION OF LAW

The use of the rules from statutes subsists in their application to specific factual circumstances. This is often much less straightforward than it would seem at the first sight. One of the chief contributors to the difficulty is the inherent imprecision of language. When the application of a general rule is not straightforward, a lawyer must present arguments as to why a provision should be applied in a particular way. In doing so the lawyer must often defend a specific account of the meaning of one or more terms. The persuasiveness and validity of a complex argument may hinge on a particular account of the meaning. Argumentation about the meaning of a term may even be the crux of an overall argument.

The level of scrutiny a communication of general legal rules is often subjected to differs radically from the case of everyday communication. It would be bizarre to respond with a series of following questions to the example utterance from Section 2.2:

> What is meant by being "cold?"
>
> What is meant by being "cold outside?"
>
> Do shoes qualify as "clothes?"
>
> Does a person "wear" a sweater if it is fasten to his or her waist?

Yet, this is exactly the kind of scrutiny general legal rules undergo regularly in the course of being applied to specific factual circumstances. As an example consider the following rule posted at the entrance to a park:

> No vehicles in the park.[1]

[1] The example is an adaptation of the rule from the classic 1958 Hart-Fuller debate over the interpretation of rules.

Focusing on the term "vehicle" there could be little doubt as to whether a car, a bus, or a motorcycle are "vehicles." It is much more challenging to decide about in-line skates, or a bike. This means that there are objects which are clearly prohibited from entering the park and no doubt with respect to this can be entertained by any reasonable person. But there are also objects in case of which it cannot be easily, if at all, determined whether they are prohibited (borderline cases). Interestingly, it should not strike anyone as bizarre, when thinking about the rule, to ask questions like these:

What is meant by "vehicle?"

What is meant by "the park?"

What does it mean to be "in" the park?

Unlike most of the terms in everyday communications, the terms from general legal rules, embodied in statutes, are expected to be examined as to their exact meaning.

In this work it is assumed that terms used in legal rules could often be imprecise as described above. This has serious implications for the examination of the meaning of terms. A theory of judicial decision making, which is sometimes called the 'standard view of adjudication,' presents the view that the judge's task is to simply give effect to the legal rights and duties of the parties. [28, p. 1] In my opinion, this view is a flagrant case of hypocrisy, hopeless idealism, or naivety. Such a view does not allow for imprecision as described above because it assumes the meaning of terms is determined—perhaps it is not immediately obvious but it could be discovered through careful analysis. The view is usually challenged with the "indeterminacy claim" which states that the requirements of the law in particular cases are frequently indeterminate. [28, pp. 1–p] This is a much more realistic view and it allows for imprecise terms.

Let us consider an example of a cyclist entering a park, who accidentally rides into a jogger. As a result both persons are injured. Although it is not clear who caused the situation, the jogger claims that bicycles are not allowed in the park. Because the cyclist was in the wrong, the jogger asks him to pay all the medical expenses. The cyclist does not agree that bicycles are forbidden from entering the park and believes that each should pay for his own expenses. If the jogger's claim is brought in front of a court the investigations

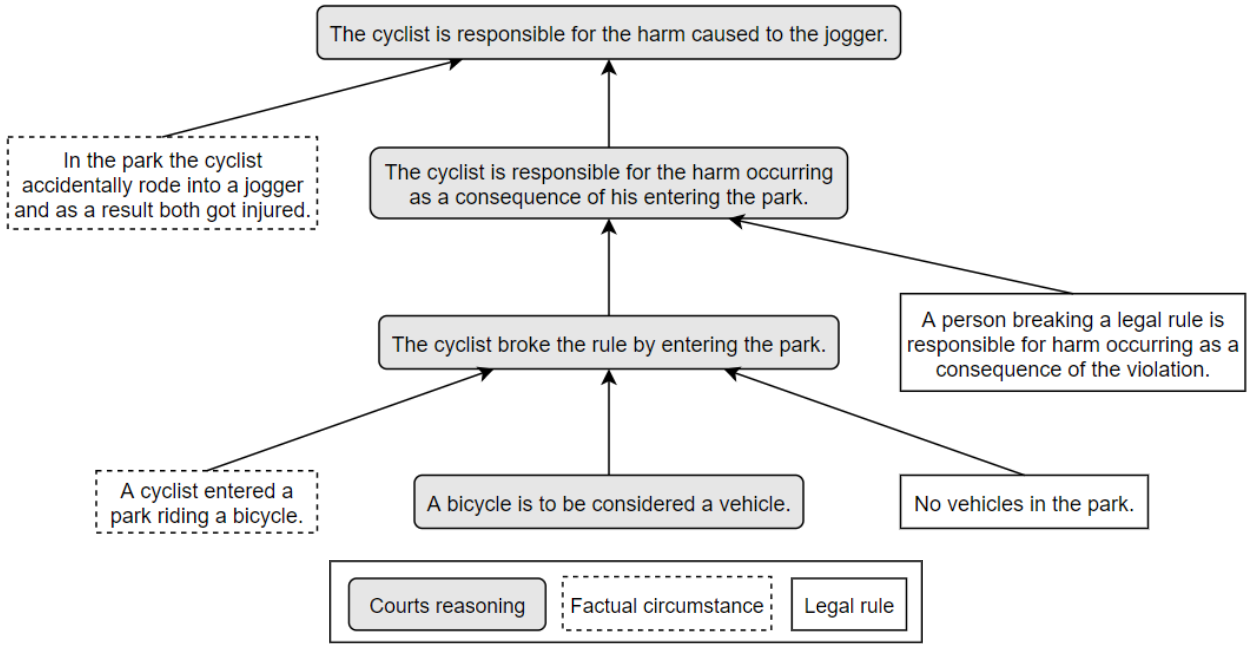

Figure 3: The Figure shows an example argument diagram where the legal responsibility of the cyclist depends on the prevailing account of the meaning of the term 'vehicle' from the example provision.

into the meaning of the term 'vehicle' from the example rule would play a central role. If such an interpretation where a bicycle is to be considered a 'vehicle' prevails, the jogger wins and the cyclist should pay for the expenses. The example argument is shown in Figure 3. An interpretation excluding bikes from the 'vehicle' category would have an opposite effect. Swapping the middle root node of the argument in Figure 3 with its negation would not allow the chain of reasoning to be instantiated. The final conclusion of the cyclist's responsibility would not be possible to reach. That is why the result of the case depends on the prevailing account of the meaning of the term 'vehicle.'

In the presented example the judge could decide either way depending on whether he or she concludes a bike is to be considered a 'vehicle' for the purpose of the rule forbidding vehicles from entering the park. The discretion implies considerable uncertainty in application of legal rules. In Hart's opinion this uncertainty is the price to be paid for the use of

general classifying terms in any form of communication concerning matters of fact. And it is indeed the price that must be paid because human legislators cannot have knowledge of all the possible combinations of circumstances which the future may bring. [44, pp. 126–128] Hence, it appears desirable to allow judges to respond to the needs of society.

For a more realistic example, let us consider the two emphasized phrases from the following (abridged) provision:

> "Enterprise" means the *related activities* performed [...] for a *common business purpose* [...].
>
> (29 U.S.C. § 203)

An understanding of the provision depends, among other things, on a well-developed knowledge of the meaning of the two emphasized phrases. The meaning of the phrase *common business purpose* from the example provision could play a pivotal role in case of, for instance, determining if two restaurants in different parts of the same city, sharing a single owner, constitute an "enterprise" within the meaning of the provision. One could for example ask if a common ownership implies a *common business purpose*.

2.4 CONSISTENT APPLICATION OF LAW AND LEGAL CERTAINTY

The cases where it is not clear if a specific legal rule applies and what effects should it have are common. This is not a flaw in legal regulation but its inherent feature. [44, pp. 124-135] Niklas Luhmann claims that the main function of law is to secure certain expectations of individuals as stable over time. [69, pp. 147–148] This relates the main function of law to certainty concerning legal rules. Legal certainty is an important value that has been traditionally recognized as crucial to the rule of law. Therefore, any threat to legal certainty should be taken very seriously. In Section 2.3 it was shown that language's imprecision (Section 2.2) gives rise to indeterminacy, hence uncertainty, in application of legal rules. As imprecision is an inherent feature of language, indeterminacy is an inherent feature of application of legal rules.

The future is uncertain. But a person wants to be certain about the future because operations in society take time. [69, pp. 143–146] For example, there may be an individual who could immediately use a fixed sum of money to generate a profit of 10%. There may be another individual who has the sum. It would be desirable if the second individual (creditor) could temporarily transfer the money to the first individual (debtor) and then both of them could share the profit. This could be done only if there is a guarantee that the debtor is going to return the money (with the agreed share of profit) to the creditor. It is the main goal of law to provide such a guarantee. [106]

Legal norms can be understood as structure of symbolically generalized expectations. By stabilized usage of this symbolization, society produces specific stabilities and sensibilities. [69, pp. 142–146] In case of the above described example the creditor can temporarily transfer the possession of the money to the debtor. It is the case because he can reliably expect that society will acknowledge his entitlement to get the money back with the agreed upon share of profit. If necessary society will help him in enforcing the legitimate claim. [106] Law makes it possible to know which expectations will meet with social approval. Given this certainty of expectations one can take on the disappointments of everyday life with a higher degree of composure. This means that one can live in a more complex society. [69, pp. 147–148]

Consider an example where the creditor lends the money to the debtor. They both agree that after a fixed period of time the money will be returned together with the half of the profit. When the time comes the debtor refuses to provide both, the money he borrowed as well as the profit he promised to deliver. Since the entitlement of the creditor to receive the money and the profit is acknowledged by society (enforceable by law) he can turn to society for help. This would usually mean that he can file a claim with a court of law. [106]

How the court addresses the claim is of vital importance with respect to legal certainty. Brian Bix offers an interesting example of a judge deciding cases on the basis of a coin-flip. [7, p. 106] One could imagine how much trust in law would be generated if the court dismisses the creditor's claim on the basis of a coin flip. Luhmann claims that where law is no longer respected, or is no longer enforced as far as it is possible so to do, the consequences extend much further than what amounts to breach of law. The system has to retreat to much more basic forms of securing confidence. [69, pp. 148] If law fails to provide members of the society

with a sufficient amount of legal certainty, i.e., fails to persuade them that their legitimate expectations will be acknowledged, it fails to perform its function altogether. [106]

The example with a coin-flipping is extreme. Such procedure would be immediately recognized as unacceptable. However, there can be more subtle forms of coin-flipping that are more difficult to recognize. One of them is closely connected with language's imprecision and resulting indeterminacy in application of law. If the indeterminacy is misunderstood in a way that in certain cases it gives a judge total freedom to decide a case that appears to be unclear, a kind of coin-flipping is being introduced in law. As has been argued above, this may have much more serious consequences than breaches of law and dismissal of legitimate claims in individual cases. [106]

2.5 PAST TREATMENT OF TERMS AND ITS ROLE IN ARGUMENTATION ABOUT THEIR MEANING

In Section 2.3 it was explained that in many cases there is room for competing interpretations of statutory terms. In the example of the claim against the cyclist causing harm to the jogger, the outcome of the case (who pays medical expenses) depended on the interpretation of the term 'vehicle.' Although conflicting, both conclusions seemed possible. Section 2.4 explained why the existence of the room for competing interpretations should not be understood as a blank permission to pick whichever understanding of the term one might prefer. This section elaborates on how do past mentions and applications of terms function as constraints on the freedom to choose an interpretation.

Cases like the example one do not exist in a vacuum. Typically there would be a number of similar cases that appeared before the present case. For the sake of clarity let us develop a scenario with two subsequent cases involving cyclists in the park. These two share a common issue of whether bicycles are allowed in the park. Apart from the facts of each case the basis for a decisions is the rule:

No vehicles in the park.

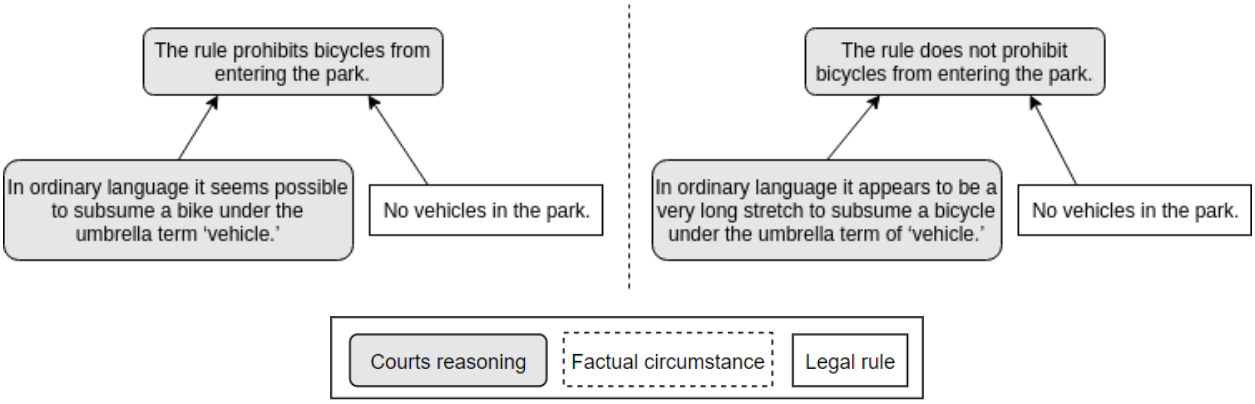

Figure 4: The Figure shows two alternative arguments leading to opposite outcomes. Interestingly, it does not appear that any of the arguments is preferable. The outcome depends on the understanding of the term 'vehicle' from the example provision.

When the first case appears in front of the court the outcome depends on the meaning of the term 'vehicle' from the example rule. In such a case the decision could probably go either way (both example arguments are shown in Figure 4). As to the two competing arguments from Figure 4 it is not possible to judge any of them as more persuasive.

In more realistic settings there would typically be a number of clues one could use for interpretation. Let us suppose that the rule was enacted after repeated complaints by the public. The merit of those complaints was that vehicles cause a lot of noise which has negative impact on the possibility to spend a pleasant time in the park. A written document related to the enactment of the rule states, among other things, that:

> The goal of the rule is to secure serenity in the park.

Furthermore, let us suppose that the cyclist was behaving very noisily to the point where he was certainly disturbing serenity in the park. Taking these into consideration the court could still decide either way. However, as shown in Figure 5 it is now easier to conclude that a bicycle should be considered 'vehicle.' Since the cyclist was clearly interfering with the goal of the rule it appears reasonable to conclude that the rule was meant to apply, among other things, to bicycles. Hence, bicycle would be deemed 'vehicle' for the purpose of the

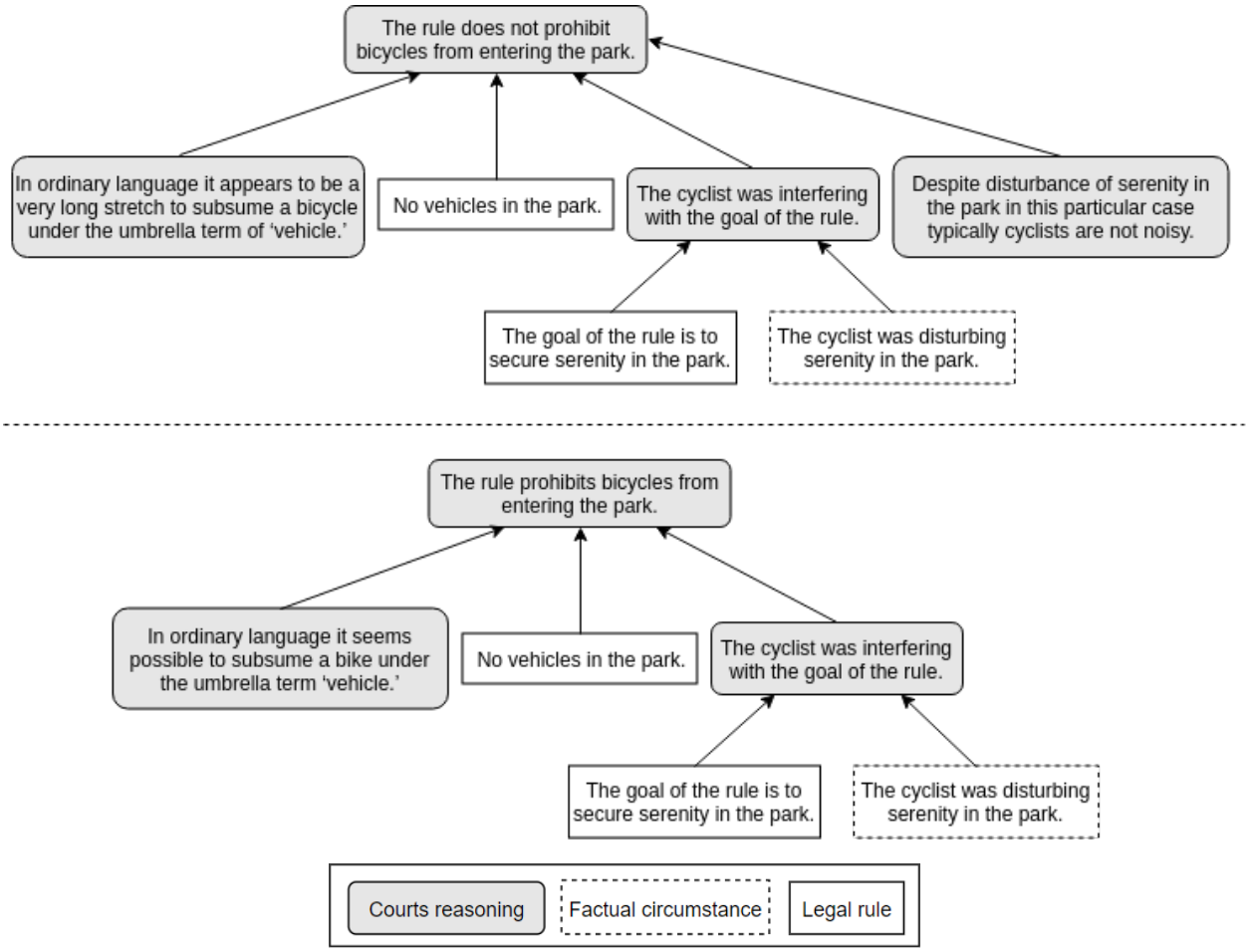

Figure 5: The Figure shows two alternative arguments leading to opposite outcomes. Here, the goal of the rule is taken into account and in light of the factual circumstances the bottom argument appears to be preferable.

rule. On the other hand, it would be more difficult to argue that the rule does not apply to bicycles since the cyclist was clearly interfering with the goal of the rule.

The above reveals an important phenomenon pertaining to the interpretation of statutory terms. The interpretation takes into account considerations that should seemingly have no effect on the meaning of the term. Normally, one would not attempt to define the term 'vehicle' with respect to disturbing serenity in the park. Yet, in statutory interpretation such a procedure is common and often desirable. One can understand why this is the case

by comparing the arguments from Figure 4 and Figure 5. Setting aside a preference for a particular conclusion, it is reasonable to assume that most of the people would agree that the arguments in Figure 5 are superior to the arguments in Figure 4. In case of the arguments concluding that the rule prohibits bicycles from entering the park, the one from Figure 5 offers an extra reason why should a bicycle be considered 'vehicle' for the purpose of the rule. In comparison, the argument from Figure 4 looks arbitrary. As to the arguments permitting bicycles, the argument from Figure 4 does little to promote the formation of legitimate expectations. A citizen should be puzzled that there is an object that could be considered 'vehicle' and that is disturbing the serenity while not being forbidden from entering the park. The argument from Figure 5 does not have this weakness.

When the second case is being decided the formation of the legitimate expectations would be best served if the previous case is taken into account. Not only is it important to decide like cases alike but it might be important to explicitly acknowledge the existence of the previous decision. Let us assume that the first case was decided in such a way that a bike is to be considered a 'vehicle' using the argument from Figure 5. In addition, let us assume that in the second case the cyclist was behaving in an orderly way and the serenity was not disturbed.

The example in Figure 6 shows how taking the previous decision into account makes the argument persuasive. Most importantly the new decision emphasizes:

1. that it is in conformance with the existing decision in the similar issue; and
2. that it treats the term 'vehicle' in the same way as the existing decision.

Such references are indispensable for formation of the legitimate expectations. It is also important that the argument deals with the fact that in the second case the cyclist was not disturbing serenity in the park. Otherwise, it could be rather confusing that there is an object which is not causing any disturbance and yet it is forbidden to enter the park. This is especially important since the disturbance was an important part of the argument in the previous case as shown in Figure 5.

Even when the new decision differs from the previous one it is still important to take it into account. Figure 7 shows that in such a situation it is important to explain why does

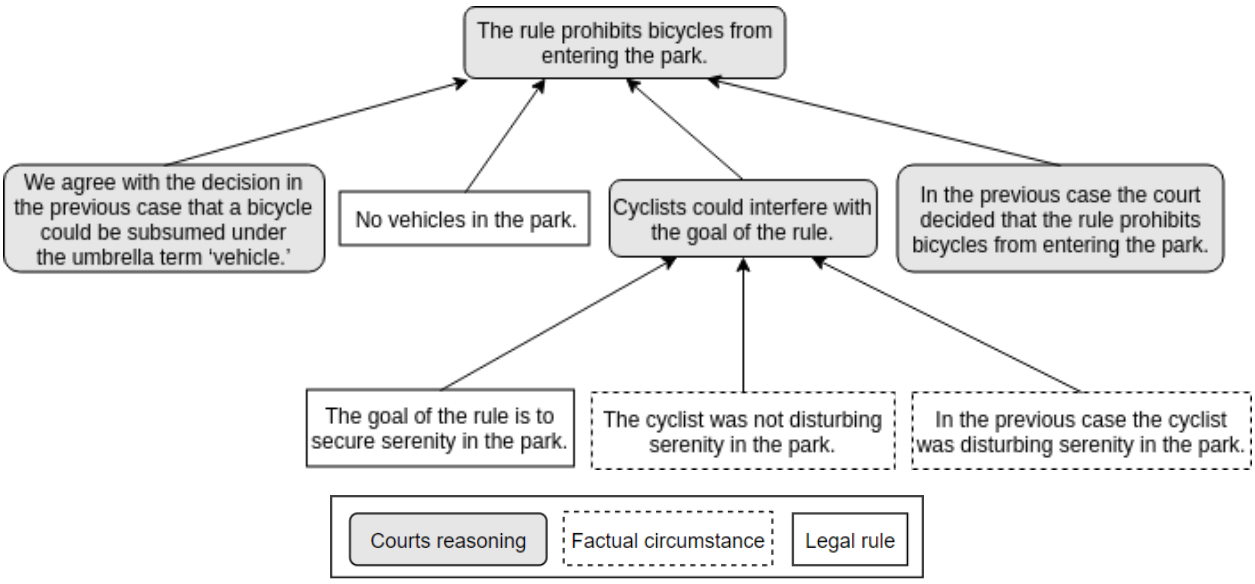

Figure 6: The Figure shows an example argument that takes into account a prior decision to which it conforms.

the new decision differ from the previous one. One possible line of reasoning, the one shown in Figure 7, could be that in the previous case the court got carried away by the fact that the cyclist was disturbing serenity in the park whereas typically cyclists would not cause any disturbance. As a consequence bicycles should not be prohibited from entering the park. Alternatively, the court could argue that a bicycle should be considered a 'vehicle' only if it causes disturbance. With respect to the formation of legitimate expectations the argument shown in Figure 7 is preferable.

I have offered a simple example of a rule being applied to two consequent cases pertaining to the same issue of whether bicycles are prohibited from entering the park. In the first case the cyclist was disturbing serenity in the park thereby interfering with the goal of the rule. In the second case no disturbance occurred. The space of possible outcomes (O1–O4) is schematically depicted in Table 1. The outcomes are generated in a process that should promote legitimate expectations and legal certainty. From this perspective outcome O2 seems clearly the least desirable one. In this case the orderly cyclist's bicycle would be

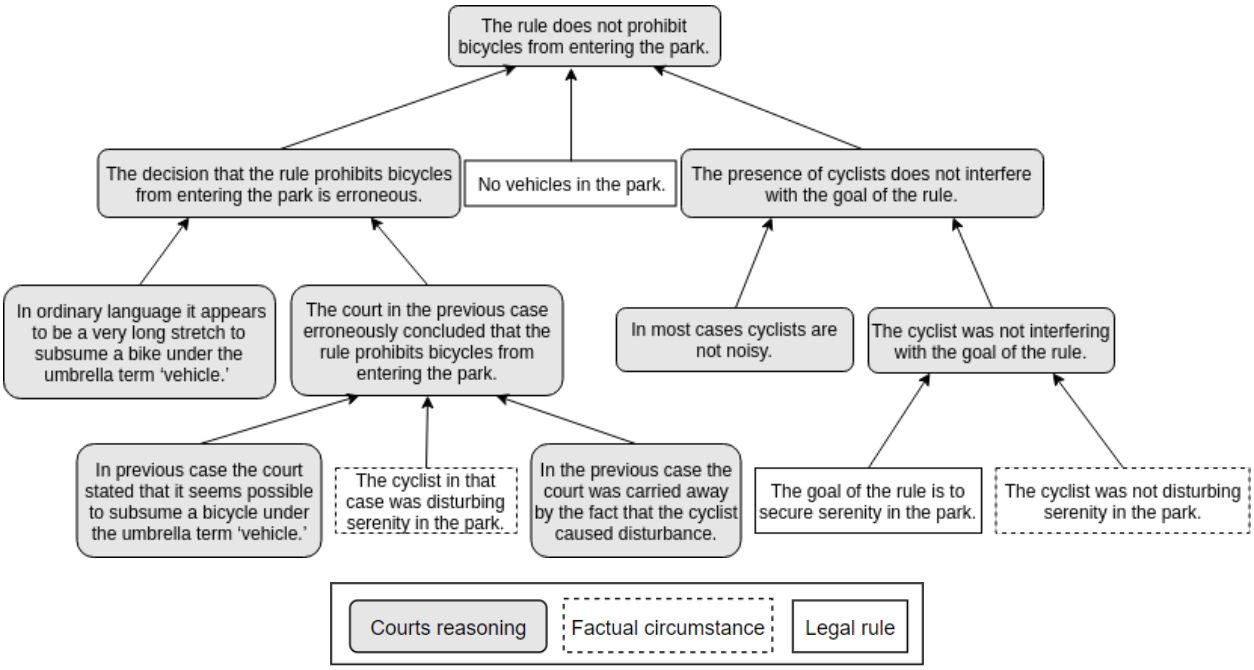

Figure 7: The Figure shows an example argument that takes into account a prior decision from which it distinguishes the case in consideration. It is important that the court acknowledges the existence of prior case law even though it decides the other way around.

considered to fall under the 'vehicle' category for the purpose of the rule promoting serenity in the park. At the same time the disturbing cyclist's bicycle would not be considered a 'vehicle' for the purpose of the rule. Given the rule, its goal, and facts of the two cases O2 would completely fail to generate any legal certainty and the process would appear arbitrary.

The outcomes O1 and O4 have the advantage that they conform to each other. In light of the example situation both outcomes would promote legal certainty. A clear message as to whether bicycles are prohibited from the park would be communicated. Interestingly, the outcome O3 is not completely undesirable. Figures 5 and 7 show the arguments that could be used in the process generating the outcome O3. However, this is not to say that O3 would be the most desirable out of the four outcomes. While the outcome O2 is clearly the least preferable one, it is not possible to decide among the other three outcomes on the basis of the available information.

		Case 1	
		Prohibited	Permitted
Case 2	Prohibited	O1	O2
	Permitted	O3	O4

Table 1: The Table shows a schematic depiction of the space of possible outcomes of the two consequent cases.

The crucial point is that difficulties in formation of legitimate expectations have very damaging effect on the legal system (Section 2.4). Therefore the outcome that best promotes the formation of legitimate expectations should be preferred. The process that generates the outcome takes into account the existing pieces of information and creates new ones. When the first case had to be decided, apart from its text, there was only the rule itself and the proclamation of its goal. Assuming the case was decided as shown in the bottom part of Figure 5 the following statements were created:

1. In ordinary language it seems possible to subsume a bicycle under the umbrella term 'vehicle.'
2. The cyclist was interfering with the goal of the rule.
3. The rule prohibits bicycles from entering the park.

When the second case was decided as shown in Figure 7 additional statements were created:

4. In the previous case the court was carried away by the fact that the cyclist caused disturbance.
5. The court in the previous case erroneously concluded that the rule prohibits bicycles from entering the park.
6. In ordinary language it appears to be a very long stretch to subsume a bicycle under the umbrella term 'vehicle.'
7. The decision that the rule prohibits bicycles from entering the park is erroneous.
8. In most cases cyclists are not noisy.
9. The cyclist was not interfering with the goal of the rule.
10. The presence of cyclists does not interfere with the goal of the rule.
11. The rule does not prohibit bicycles from entering the park.

The statements may apply general rules to specific cases (e.g., the statements 3. and 9.), propose new rules (e.g., the statement 10.), further specify existing rules (e.g., the statements 3. and 11.), comment on existing propositions (e.g., the statements 5. and 7.), or elaborate on the meaning of statutory terms (e.g., the statements 1. and 6.). Once pronounced these elaborations become part of the whole of past treatment of a statutory term.

2.6 ANALYSIS OF PAST TREATMENT OF TERMS

Understanding of a statutory provision depends on well-developed knowledge of the meaning of its constituent terms and phrases. Doubts about the meaning of a term may be removed by interpretation (Sections 2.1 and 2.3). The interpretation involves an investigation of how a term has been referred to, explained, interpreted, or applied in the past (Section 2.5). The analysis of such past treatment is an important step that enables one to then construct arguments in support of or against particular interpretations. A particular account of the meaning of the term of interest is scrutinized against the results of the analysis. Specifically, it is important to show that the account of the meaning makes sense in the context of the past treatment.

The past treatment consists of text passages that mention or use the term of interest. These passages may come from many different sources. Typically, they come from past court decisions, journal articles, conference papers, or legislative histories. In this work I focus on passages coming from court decisions, the most important of the sources. A court decision is a result of court proceedings that gives rise to, modifies, annuls, or acknowledges specific legal rights and obligations of persons, organizations, businesses, etc. Court decisions are an important source of legal information in both, common law and continental legal systems. Court decisions contain information about how the courts applied generally binding legal rules to the actual cases in the past. Court decisions are probably the most valuable source of information about the actual impact of statutory law.

The elaboration on the meaning of legal rules from statutes is not the only type of information that can be found in court decisions. A court decision could from a certain

perspective be viewed as a solution to a specific legal problem including reasoning that led to the solution. Often a court must assemble and consider a vast amount of different types of information and organize them in such a way that the solution follows from the available information. This information may include but is not limited to statements of legal rules, court's conclusion as to whether a legal rule's requirement has been satisfied, fact trier's findings of fact, elaboration on the merits of a case, such as legal factors, stereotypical patterns of fact that strengthen or weaken a side's legal claim, and arguments a judge or the litigants are making. [5]

A lawyer that needs to argue about a meaning of a term will most likely use some of the available commercial legal IR systems to analyze the past treatment. The most likely source of useful passages are court decisions. Thus, a lawyer would search through the database of court decisions and inspect mentions and uses of the term of interest. Although, this strategy works it is labor intensive and often not very effective. A lawyer may need to go through many decisions before putting together a reasonable set of useful sentences or before concluding that such a set most likely does not exist.

A lawyer analyzing the past treatment of the term 'vehicle' from the example rule prohibiting vehicles from entering the park would retrieve the two decisions (the one from the bottom part of Figure 5 and Figure 7) with the following passages mentioning the term:

The decision from the bottom part of Figure 5

> No *vehicles* in the park.
>
> In ordinary language it seems possible to subsume a bike under the umbrella term '*vehicle*.'

The decision from Figure 7

> No *vehicles* in the park.
>
> In previous case the court stated that it seems possible to subsume a bicycle under the umbrella term '*vehicle*.'
>
> In ordinary language it appears to be a very long stretch to subsume a bike under the umbrella term 'vehicle.'

Even from the toy example it is apparent that the lawyer will have to sift through decisions with passages that are often useless (e.g., verbatim citation of the rule) and redundant.

A realistic example would involve anything from hundreds to tens of thousands of decisions. The decisions would be ordered per some standard IR relevance measure. This

roughly means that the decisions having an unusually high occurrence of the term of interest would appear at the top of the list. This is a useful approximation since the decisions that are about the term of interest are likely to contain many mentions of the term. In most of the cases, a lawyer would be required to manually analyze at most hundreds of documents before the analysis would be deemed complete. Such a process is still very expensive and contains many inefficiencies, such as the following:

- An analyzed court decision would likely contain many mentions of the term of interest. Typically, only a small number of these mentions would be useful for argumentation about the meaning of the term of interest. Yet, a lawyer would have no alternative but to carefully examine every single mention in a decision which has been ranked high.
- There would likely be a high redundancy of mentions across analyzed decisions. For example, one would come across many citations of the statutory provision containing the term of interest. There could also be a number of repeatedly cited court-made statements that contain the term but have little use in its interpretation. A lawyer would have no effective means to filter these out.
- There may be numerous mentions of homonymous terms, i.e., the terms that have the same spelling but different meaning. These are typically much less useful (if at all) for the past treatment analysis. An overwhelming presence of homonymous mentions may greatly contribute to the cost of the analysis.
- There might be an important mention appearing in a longer decision that does not contain many other mentions of the term of interest. In such a case the decision might be very low in the results list. A lawyer would have to analyze all the decisions in the list to make sure such a mention is not missed. This would often be prohibitively expensive.

The lawyer that eventually arrives at a small number of mentions (perhaps, less than 5) often needs to inspect hundreds if not thousands of text passages mentioning the term of interest.

Legal encyclopedias and dictionaries could be a useful alternative resource. They typically do not suffer from the high redundancy and irrelevancy problems. However, since these are being hand-crafted by legal experts they are very expensive to create and maintain. In

addition, their contents are limited. The commercial product that is the closest in functionality to the subject of this work is the Thomson Reuters' Words and Phrases®. The electronic product is based on the printed counterpart. It lists judicial definitions, from both state and federal courts, of words and phrases, arranged alphabetically. Definitions may pertain to statutory language, court rules, administrative regulations, or business documents, among other sources. [1] It appears that there is a significant human expert contribution to the identification of definitions.

3.0 TASK

3.1 TASK'S CONTEXT

The task that this work automates is the retrieval of a statutory term's past treatment from case law. Typically, this task is performed in a wider context of solving a legal problem. Taking the example of a cyclist and a jogger colliding in the park the problem is the occurrence of medical expenses and the jogger's belief that it is not fair for him to bear them. The solution to the problem would be identifying someone else to bear the costs. As one is thinking about possible strategies of reaching the solution he or she is gradually framing the problem in terms of one or more issues that need to be solved. Alongside the issues, one also develops sets of potentially relevant rules and facts (left part of Figure 8).

At the end of the initial stage one understands what issues need to be resolved through application of general legal rules onto the specific facts of the case. In the example the key issue could be whether the cyclist broke the rule by entering the park. We know for a fact that the cyclist entered the park riding a bicycle. As for the applicable rules we know that the rule states: 'No vehicles in the park.' Interestingly, it should be quite obvious that these two pieces of information are not sufficient basis for resolving the issue. Such a situation where there are doubts as to if and how does a rule apply to the facts at hand is common. That is why interpretation is an integral part of the application as shown in Figure 8. The process is iterative in a sense that as one applies the rules to the facts doubts may arise and as they are removed the application results could change which leads to a new set of doubts.

The doubts may arise due to several causes where the two most prevalent ones are:

1. the use of imprecise language as described in Sections 2.2 and 2.3

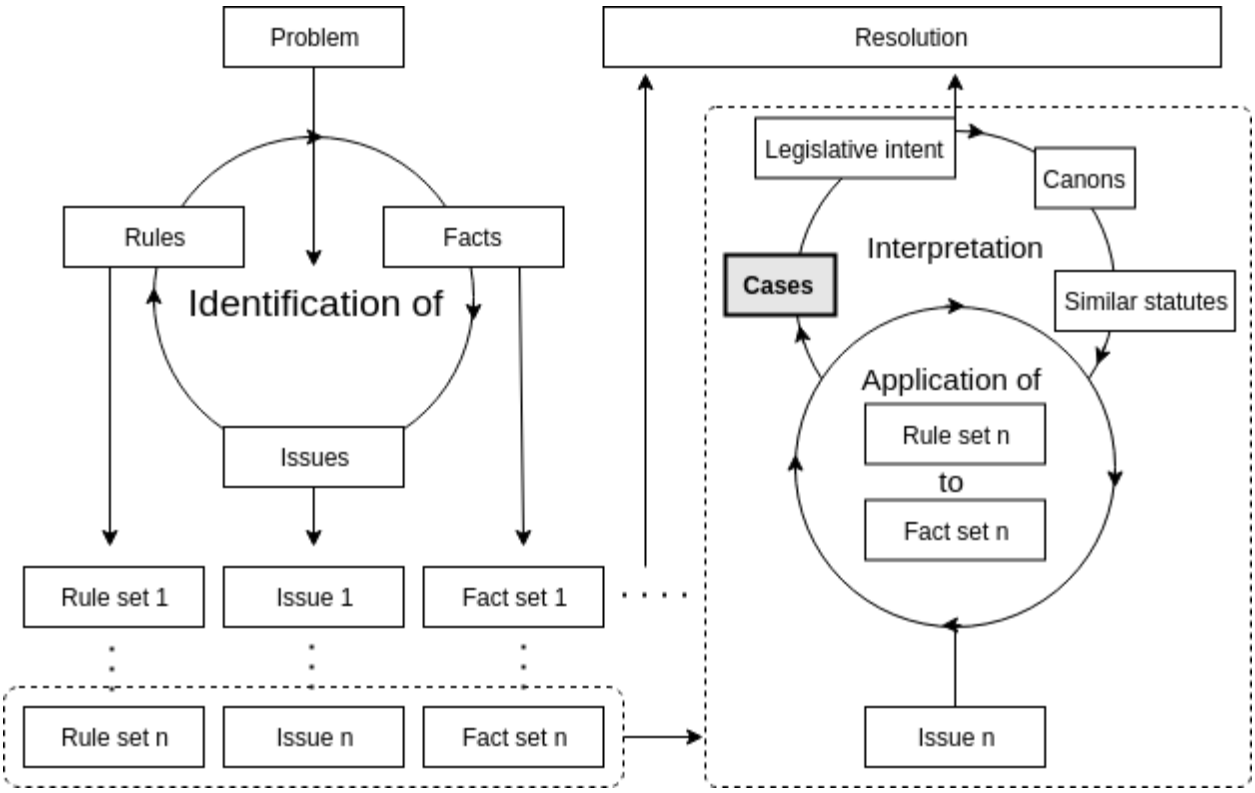

Figure 8: The figure depicts the schema of a process of legal problem solving. This work focuses on the support of the past treatment analysis which takes place at the 'Cases' node in the figure.

2. the silence of the rule with respect to some aspect of the issue

The doubts are removed by interpretation. This work focuses on the support of interpretation in case of the use of imprecise language. In Figure 3 the interpretation consists of a simple statement that: 'A bicycle is to be considered a vehicle.' This statement adds the missing piece necessary for a resolution. Typically, the situation is not so simple and one is not allowed to arrive at such a statement without following an established procedure. The authors of [109] explain that statutes are like the rules of a game. When the rules of the game are unclear, there are several possible means of clarification:

1. One can puzzle through the language (plain meaning rule, purpose approach, golden

rule, canons of construction).

2. One can seek an authoritative interpretation (past treatment in case law).
3. One can try to discern the point of the game and interpret the unclear rule accordingly (legislative intent).
4. One can draw on his knowledge of the rules of other similar games (similar statutes). [109]

The process is schematically depicted in the right part of Figure 8. This work focuses on the support of the past treatment analysis which takes place at the 'Cases' node in the figure.

In the context of the example resolution of the issue shown in Figure 7 the analysis of past treatment identifies the existing mention of the term of interest in the sentence: 'It seems possible to subsume a bicycle under the umbrella term vehicle.' The mention points one to the relevant case which then drives the reasoning in the more important left branch of the argument. While the construction of the argument is out of the scope of this work we focus on the retrieval of the relevant mentions.

3.2 SOURCE PROVISION AND THE TERM OF INTEREST

As explained in Section 2.1, statutory provisions express the individual legal rules (e.g., rights, prohibitions, duties). A provision consists of one or more sentences that together express a single rule. The definition is fluid and in fact a provision is not a clearly identifiable structural unit such as a chapter or a section (see Section 2.1). While sometimes a whole list of definitions could be referred to as a provision, other times each definition from the list would be called a provision of its own. In this work I subscribe to the latter and as a provision I treat the smallest possible piece of statutory text that fully expresses a single rule, goal, definition, etc.

For the purpose of this work the *source provision* is a provision which is the focus of interest and needs to be interpreted. In the source provision I identify a single word or multi-word phrase that might be imprecise and refer to it as the *term of interest*. Figure 9 shows an example source provision, definition of 'wire communication,' with 'aural transfer' as an example term of interest (in bold). It should be emphasized that there might be terms

TITLE 18. *Crimes and Criminal Procedure* — PART I. *Crimes* — CHAPTER 119. *Wire and Electronic Communications Interception and Interception of Oral Communications*
18 U.S. Code § 2510(1)
(1) "wire communication" means any **aural transfer** made in whole or in part through the use of facilities for the transmission of communications by the aid of wire, cable, or other like connection between the point of origin and the point of reception (including the use of such connection in a switching station) furnished or operated by any person engaged in providing or operating such facilities for the transmission of interstate or foreign communications or communications affecting interstate or foreign commerce;

Figure 9: The figure shows an example source provision, definition of 'wire communication,' with 'aural transfer' as an example term of interest (in bold).

identical to the term of interest appearing in other provisions (i.e., 'aural transfer' mentioned in a provision other then 18 U.S. Code § 2510(1)). These may or may not have the same meaning. The difference in meaning of terms coming from different statutory provisions may be more pronounced than in terms used in everyday communication. This is due to a unique way in which statutory terms acquire meaning through interpretation in the context of specific cases (Figures 3–7). A typical example of terms having the same meaning would be a term defined in a Definitions section of a chapter and the same term being mentioned elsewhere in the chapter (e.g., 'wire communication' in 18 U.S. Code § 2510(1) and in 18 U.S. Code § 2511). General terms used in different legal domains are clear examples of terms with different meaning (e.g., 'work' in employment law and copyright law). The sameness of the meaning is one of the key factors in determining the relevance of a mention (see Section 3.4).

3.3 CASE LAW SENTENCES MENTIONING THE TERM OF INTEREST

As explained in Section 3.1 this work focuses on supporting the term of interest's (see Section 3.2) past treatment analysis in case law. The goal is to retrieve text passages that use or mention the term of interest. Importantly, only those passages that are useful for

interpretation of the term of interest need to be retrieved. In Section 3.4 we explain how to determine the utility of a sentence. Here we focus on the concept of a 'text passage.'

As explained in Section 2.6 the retrieval of useful mentions of the term of interest typically starts with a list of the whole decisions that a lawyer then needs to manually inspect. Our approach differs in that we retrieve the passages directly in hope of mitigating the issues described in Section 2.6. Since decisions of the courts are traditionally recognized as the most prominent source of the useful mentions we retrieve the passages from case law. The important question we needed to answer was as to how do we define a passage that mentions the term of interest. Essentially, there were two unknowns:

1. What does it mean for a passage to mention the term of interest.
2. Where does a text passage start and where does it end, i.e., what are the passage's boundaries.

With respect to the first issue I opted for a very simple solution of requiring a passage to actually contain at least one exact match of the term of interest. By exact match I mean that for a phrase we require the identical sequence of word lemmas to be present in the passage. There are two linguistic phenomena that are potentially threatening to make this approach fall short—synonymy and co-reference. I assume that synonymy is likely a very minor issue. Courts tend to use very precise language in their opinions. It does not appear very likely that in case of a passage dedicated to the interpretation of a specific term of interest coming from a specific source provision, a court would use a synonymous expression in place of the term of interest.

Co-reference is a real issue and not taking it into account is a limitation of this work. Consider the two example passages that follow in a sequence with respect to 'vehicle' as the term of interest:

> The term 'vehicle' is very general.
>
> Therefore, in ordinary language it seems possible to subsume even a bicycle under its umbrella.

Intuitively, both of the passages seem useful whereas only the first one could be retrieved using our approach. Dealing with this issue is a difficult problem and taking it into account

would make numerous stages of the work significantly more complicated. Especially the initial stage of assembling the data set would be difficult. First, one would have to manually go through the full decisions and make sure that any co-referent to the term of interest is identified. Second, an annotation process would have to be altered in order to make sure that the sentences that do not contain the term of interest itself would be presented in a meaningful way. I decided to ignore the issue for now. It appears to be relatively self-contained and I expect it would be possible to extend the work with the techniques to deal with the issue in future work. At the moment I do not understand this work as a full automation of the past treatment retrieval. Instead, I understand the retrieved passages as pointers, the original context of which should be manually investigated in order to complete the analysis.

With respect to the second issue (boundaries of a passage) there are multiple approaches that could have been taken. I considered the three following options:

1. A fixed width *window* consisting of n tokens on each side of the term of interest.
2. A complete *sentence* containing the term of interest.
3. A complete *paragraph* containing the term of interest.

The advantage of a fixed width window would be its simplicity. On the other hand, determining the right size for the window would be a challenge. The paragraph, defined as a consecutive stream of tokens between two line breaks, is very appealing. First, it is very simple to segment full decision texts into paragraphs. Second, a paragraph typically serves as a container for a closely related stream of thoughts. Using paragraphs would have the advantage of presenting the user with a relatively rich self-contained context for each of the term of interest's mention. The main disadvantage is that paragraphs are most likely too large and a single paragraph would most likely contain multiple mentions of a term of interest. In fact, a decision that discusses the meaning of a term of interest would most likely do so in one or just a few paragraphs. Yet, a majority of mentions in that paragraph would most likely not be useful (see Figure 10 for an example). Therefore, it would seem that using a paragraph as a passage would deprive our work of a large portion of its value.

I chose the sentence as the most appropriate passage container. A sentence, as a se-

> During the hearing on the Defendant's motion to dismiss the information on the grounds that section 790.225 was unconstitutionally vague, the trial court examined the **switchblade knives** confiscated from the Defendant as well as information provided by the Defendant about a Russian-made ballistic knife that shoots knife blades. The trial court also reasoned that the Legislature did not intend to prohibit the sale or possession of **switchblade knives** and that to uphold the constitutionality of the statute "would be to make the **switchblade knife** an illegal weapon in Florida." When the statute is read as a whole, the statute imparts to a person of common intelligence and understanding that possessing a **switchblade knife** would be a crime. Using such terms, then, it seems apparent the Legislature intended to make the possession of **switchblade knifes**, which "*propel a knife-like blade as a projectile by means of a coil spring, elastic material, or compressed gas*," illegal while allowing possession of other types of knives or pocketknives.

Figure 10: The figure shows an example of a passage consisting of four sentences each mentioning "switchblade knife" (bold). Yet, only the last mention has high value for argumentation about the meaning of 'switchblade knife" (emphasized citation).

quence of grammatically linked words, is capable of expressing (at least implicitly) a complete thought. [15] This is a clear advantage over the fixed window of n tokens. Unlike paragraphs, sentences are definitely not too large. If anything they could occasionally be too small where a closely related couple of sentences would be the most appropriate size for a passage. Retrieving passages consisting of one or more sentences could be a valuable extension of this work. However, this would be a non-trivial task and I leave it for future. In many cases a single sentence would be an appropriate size for a passage. The disadvantage of using a sentence as a passage is that detecting sentence boundaries in case law is an unsolved problem (see [107]).

In conclusion, a response to a query about a term of interest from a source provision has the form of a list of case law sentences having one or more exact matches to the term of interest. Despite a number of limitations discussed in this section I believe the approach to have enough value to warrant this work. For the time being, the limitations are ignored for the sake of making the work feasible.

3.4 SENTENCE'S UTILITY

Earlier I explained what role does the analysis of a past treatment of a term of interest play in the application of general legal rules to specific facts of the case in the context of legal problem solving (Section 3.1). I also explained how the retrieval of sentences (i.e., passages of text) that mention the term of interest is an important step in the analysis (Section 2.6). Finally, on multiple occasions I hinted that not all of the sentences are useful for the analysis (Section 2.6).

Consider the following (abridged) source provision (29 U.S. Code § 203) with the example (emphasized) term of interest:

> "Enterprise" means the related activities performed [. . .] for a *common business purpose* [. . .].

The meaning of the phrase *common business purpose* from the source provision could play a pivotal role in case of, for instance, determining if two restaurants in different parts of the same city, sharing a single owner, constitute an "enterprise" within the meaning of the provision. The interpretation involves an investigation of how the term has been referred to, explained, interpreted or applied in the past. This is an important step that enables a user to then construct arguments in support of or against particular interpretations.

Searching through a database of statutes, court decisions, or law review articles, one may stumble upon sentences such as these:

(1) Courts have held that a joint profit motive is insufficient to support a finding of *common business purpose*.
(2) The fact of common ownership of the two businesses clearly is not sufficient to establish a *common business purpose*.
(3) Because the activities of the two businesses are not related and there is no *common business purpose*, the question of common control is not determinative.
(4) The problems then are whether we have related activities and a *common business purpose*.
(5) The third test is "*common business purpose*."

Some of these sentences are useful for interpreting the phrase *common business purpose* (1 and 2). Some of them look like they may be useful (3) but the rest appears to have very little (4) if any (5) value. Reviewing such sentences manually is labor intensive due to high

redundancy and the large number of sentences that simply quote the statutory language but do not add any information (Section 2.6).

The goal of this work is to evaluate and propose computational methods to support this task. Specifically, given a user's interest in the meaning of a particular term of interest, it is desirable to rank more highly sentences the goal or effect of which is to elaborate upon the meaning of the term of interest. These include but are not limited to:

- definitional sentences (e.g., a sentence that provides a test for when the phrase applies),
- sentences that state explicitly in a different way what the statutory phrase means or state what it does not mean,
- sentences that provide an example, instance, or counterexample of the phrase, and
- sentences that show how a court determines whether something is such an example, instance, or counterexample.

By contrast, sentences that merely quote or closely paraphrase the statutory language should be demoted. Even though they may contain instances of the statutory phrase such as "related activities" or "common business purpose," they do not help because they contain no additional information about the meaning of the phrase.

3.5 TASK DEFINITION

In this section I define the task of discovering sentences for argumentation about the meaning of statutory terms. The query that describes the information consists of two elements:

- A *source provision* as described in Section 3.2 which is typically a moderately short text consisting of one or several sentences.
- A *term of interest* as described in Section 3.2 which is typically a phrase consisting of one or several words.

The database of case law (as described in Section 2.6) has the following constituents:

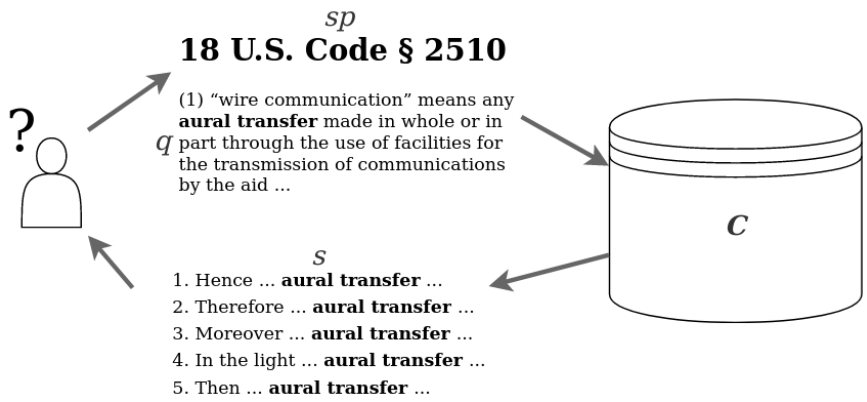

Figure 11: The figure shows the task from the user's perspective. The source provision (sp) and the term of interest (q) are submitted as a query. The system retrieves a list of sentences (s) from the database of case law (C). The sentences are presented to the user.

- *Single cases* (court decisions) as described in Section 2.6 which are variably sized texts typically consisting of several to many *sentences*.

The goal is to retrieve a list of case law sentences that mention the term of interest from the source provision ranked in such a way that those sentences that are useful for the interpretation of the term of interest (as described in 3.4) appear on top of the list.

Figure 11 shows the task from the user's perspective. The user submits the source provision with the term of interest as the query to the system which is equipped with a database of case law. In response he or she receives the ranked list of sentences coming from the individual cases. Although, this looks rather straightforward, the task does not map directly onto any well-established IR or text mining task. There is an important step that remains implicit in Figure 11—the segmentation of individual cases into sentences. This step is not only a necessary prerequisite but as it turns out, it is also a challenging task on its own (see [107]). Assuming it is possible to segment cases into sentences, there is a number of different approaches as to how the task of discovering sentences for argumentation about the meaning of statutory terms could be mapped onto some more general IR or text mining task.

The simplest approach would be to segment all cases into sentences upfront. Thus, one would end up with a database of sentences instead of the database of cases. Using this approach the task would map straightforwardly to ad hoc document retrieval. In a classical IR framework, one would then measure similarity of the term of interest to a sentence, using a measure such as BM25, [73, p. 233] and present the user with the list of best matching sentences. Another simple approach would be to first retrieve all cases that match the term of interest and rank them using again a measure such as BM25. The top n results could then be segmented into sentences. All the sentences matching the term of interest could finally be presented to the user. A conceptually similar approach is typically used in question answering where documents that likely contain an answer are retrieved first. From these a set of candidate answers (short passages of text) are then generated and ranked.

Either of the two simple approaches described above would not be very successful without further improvements. However, they present two different paradigms, each with its own set of challenges, in terms of which the task could be understood. This work takes inspiration from both of the approaches. As shown in Figure 12 the main part is the ranking component (shown as the grey triangle in the middle). The function of the ranking component is to take the full list of case law sentences mentioning the term of interest and reorder (rerank) them in a way where the most useful sentences (Section 3.4) end up on the top of the list.

It appears to be well-established that the approach to ranking based on computing similarity between the query and retrieved documents is less effective for very short documents such as sentences (see, e.g., [81]). The results of our experiments reported in Section 6.1 confirm that similarity-based matching of the term of interest to sentences themselves does not work well. That is why the ranking component needs to take into account other parts of the cases the sentences come from (i.e., sentences' context), the whole source provision (i.e., the term of interest's context), as well as other resources, such as pre-trained language models. The crux of this work is to explore how such resources could be used in service of different ranking techniques to facilitate an effective discovery of sentences for argumentation about the meaning of statutory terms.

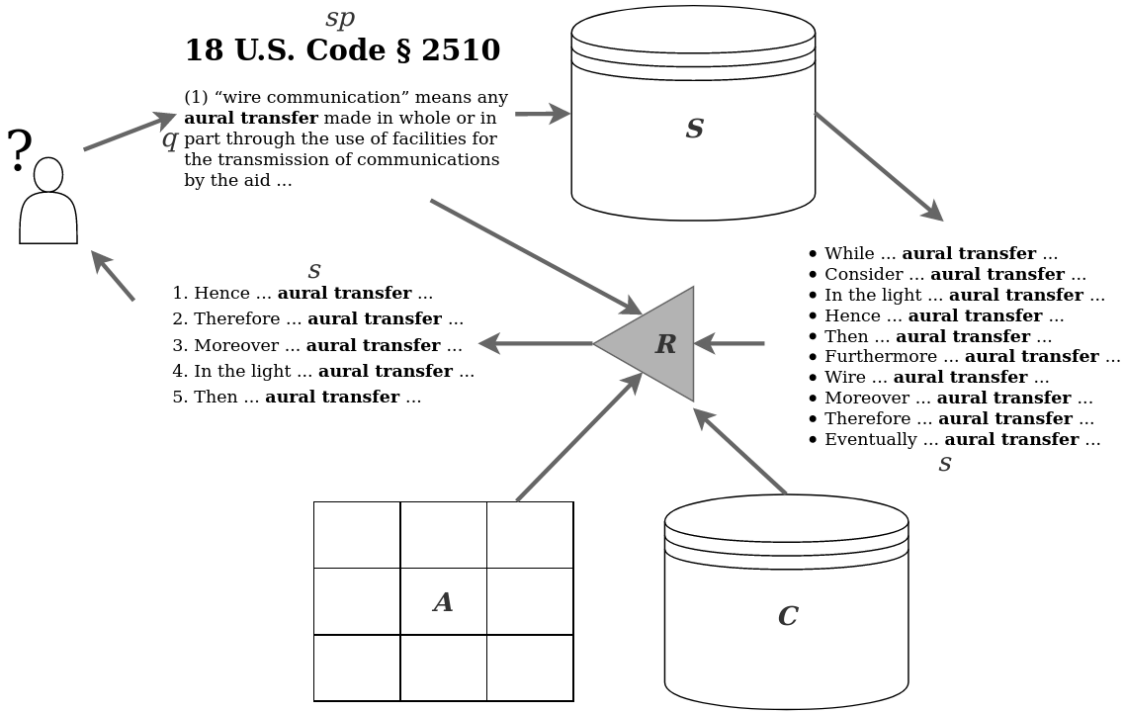

Figure 12: The Figure shows the main task components. The source provision (sp) and the term of interest (q) are submitted as a query. The system retrieves a list of sentences (s) from the database of sentences (S). The retrieved sentences are then passed through the reranking component (R) to determine which of the sentences in which order should be presented to the user. The reranking component takes into account sentences' context (full case texts from case database C), the source provision, as well as other resources, such as pre-trained language models.

4.0 DATA SET

The statutory interpretation data set was created in three major iterations between 2016 and 2019. In 2016 the task of case-law sentence retrieval for statutory interpretation was first defined. I have drafted the first version of the annotation guidelines (Appendix A). These were subsequently used to annotate a rather small set of about 250 sentences concerning three statutory terms (see [104] for details). In 2018 the data set was extended with other sentences related to the same three terms using the same annotation guidelines. The data set's size increased more than 18 times (see [108] for details). Finally, in 2019 11 law students were hired to annotate a sizable data set comprising of more than 20,000 sentences. For this final iteration a more detailed version of annotation guidelines was created (Appendix B). The annotation effort as well as the resulting corpus of almost 27,000 sentences are described in this chapter.

In [104] the corpus consisting of 243 sentences was used for the initial feasibility study. In [108] me and the co-authors created a significantly larger data set of 4,635 sentences. That data set supported our experiments in identifying or confirming some of the challenges specific to sentence retrieval, as well as problems specific to the task of retrieving case-law sentences for statutory interpretation. The data set was also sufficient for evaluation of different ranking methods with respect to which of them work well or have interesting properties. However, a data set of this size arguably did not allow:

- drawing of finer grained conclusions as to which of the methods perform better if their performance appears to be quite similar,
- truly principled setting of hyperparameters for some of the ranking methods, and
- application of larning-to-rank approaches that learn from the human created relevance

labels.

Overcoming these limitations was the main driving force behind extending the data set.

A significant extension of the data set from three to tens of queries required a considerable time commitment from the annotators. From the past experience I knew that a trained annotator is capable of producing a label per 30 seconds on average. Reaching the goal of more than 20,000 labeled sentences could require anything from 40,000 to 60,000 individual annotations. This is because each sentence needs to be seen by two annotators and then adjudicated by the third in case there is a disagreement. The whole task therefore required approximately 500 hours. Dividing this amount among the three annotators from [108] would translate into roughly a month fully dedicated to sentence labeling for each of them. Since that was not feasible external annotators were hired. On one hand, this presented me with certain challenges, the most significant of which were:

- identifying and securing a competent and productive group of annotators,
- drafting of the more detailed and explicit annotation guidelines,
- identifying a larger group of the terms of interest,
- training the annotators, and
- streamlining and overseeing the annotation process involving the larger group of external annotators.

On the other hand, the use of the annotators that are not the authors of the work has the benefit of offering unique insights into the task itself, such as how difficult and objective the task is, what does it take to train the external annotators, etc.

4.1 ANNOTATORS

To facilitate the annotation process a University of Pittsburgh Pitt Cyber Accelerator Grant entitled "Annotating Machine Learning Data for Interpreting Cyber-Crime Statutes" was secured. The grant provided resources for the external annotators' salary which was set to $15 per hour. As for the annotators themselves law students from the University of

Pittsburgh's School of Law were recruited. The choice was chiefly motivated by the three following factors:

- The rate of $15 per hour was expected to be acceptable for the students whereas it would probably not be sufficiently attractive to lawyers who already obtained their degree.
- The law students had already received part of their legal training and were expected to perform better than general population.
- We assumed that the task itself would have been appealing for the students and they would have been well motivated to perform well.

Eventually 11 law students applied for the position of a research intern and were hired as annotators.

4.2 EXTENDED ANNOTATION GUIDELINES

The original annotation guidelines (Appendix A) used during the creation of the proof of concept [104] and the second iteration of the data sets [108] were designed for the authors of the work. This means that they could leave certain aspects of the task implicit because the annotators were at the same time the designers of the task. For the hired law students it was necessary to extend the guidelines and make the implicit assumptions explicit. Therefore, I drafted the second version of the guidelines which is in Appendix B. The new version of the guidelines is compatible with the first one but contains many details that were not part of the original version. The new version also addresses the issues that were identified during the creation of the first two data sets. The sentences were classified into four categories according to their usefulness for the interpretation of the corresponding term:

1. **high value** - This category is reserved for sentences the goal of which is to elaborate on the meaning of the term. By definition, these sentences are those the user is looking for.
2. **certain value** - Sentences that provide grounds to draw some (even modest) conclusions about the meaning of the term. Some of these sentences may turn out to be very useful.

3. **potential value** - Sentences that provide additional information beyond what is known from the provision the term comes from. Most of the sentences from this category are not useful.

4. **no value** - This category is used for sentences that do not provide any additional information over what is known from the provision. By definition, these sentences are not useful for the interpretation of the term.

4.3 DATA

I downloaded the complete bulk data from the Caselaw access project.[1] This includes all official, book-published U. S. cases from all federal and state courts as well as from a number of territorial courts [90]. Altogether the data set comprises more than 6.7 million unique cases. I ingested the data set into an Elasticsearch instance.[2] For the analysis of the textual fields I used the LemmaGen Analysis plugin[3] which is a wrapper around a Java implementation[4] of the LemmaGen project.[5] The lemmatizer is based on so-called induced ripple-down rules. [58] To support the experiments I indexed the documents at multiple levels of granularity. Specifically, the documents were indexed at the level of full cases as well as segmented into the head matter and individual opinions (e.g., majority opinion, dissent, concurrence). This segmentation was performed by the Caselaw access project using a combination of human labor and automatic tools.[6] I also used the sentence segmenter from [107] to segment each case into individual sentences and indexed those as well. Finally, I used the sentences to create paragraphs. I considered a line-break between two sentences as an indication of a paragraph boundary. Including indexes the resulting data set is 622GiB in size and has nearly 0.8 billion documents. More detailed statistics are reported in Table 2.

[1] A small portion of the data set is available at `case.law`. The complete data set could be obtained upon entering into research agreement with LexisNexis.

[2] `https://www.elastic.co/`

[3] `https://github.com/vhyza/elasticsearch-analysis-lemmagen`

[4] `https://github.com/hlavki/jlemmagen`

[5] `http://lemmatise.ijs.si`

[6] Per information provided in the email from info@case.law on 2019-01-07.

cases	6,715,418
opinions	13,958,364
paragraphs	206,058,346
sentences	538,680,570

Table 2: The table shows summary statistics of the case law database used in this work.

I queried the system for sentences mentioning 39 new terms to supplement the original three used in [108]. The terms (including the original three) are listed in Table 3 with the full citation and the Title they come from. Ideally the term selection process would be random. However, this was not possible and I had to conform to the following constraints:

1. The list of results returned in response to the term as the query had to be at most 5000 sentences; and
2. The list of results had to contain at least one sentence which intuitively appeared to be of high value for the argumentation about the meaning of the term.

The first constraint ensured that excessive resources would not be spent on a single query since approximately 20,000 sentences could be labeled. The role of the second constraint was to guarantee that the resources would not be spent on terms that would have been of no use for the purposes of this work. Furthermore, it was my goal for the terms to come from different areas of law. Given these constraints I made my best effort to create a well-balanced data set.

4.4 TRAINING OF THE HUMAN ANNOTATORS

The first step in the annotation process was to train the law students. The training consisted of the seven following modules:

1. **Initial meeting** – The training started with a 60 minutes long group meeting in which

Term of Interest	Title	Citation
identifying particular	5. Government Organization and Employees	5 U.S. Code §552a(a)
security vulnerability	6. Domestic Security	6 U.S. Code §1501(17)
cybercrime	6. Domestic Security	6 U.S. Code §1531(a)
viticultural	7. Agriculture	7 U.S. Code §451
hybrid instrument	7. Agriculture	7 U.S. Code §1a(29)
accommodation trade	7. Agriculture	7 U.S. Code §6c(2)
dischargeable consumer debt	11. Bankruptcy	11 U.S. Code §722
fully amortize	12. Banks and Banking	12 U.S. Code §4901(1)
electronic signature	15. Commerce and Trade	15 U.S. Code §7006(5)
significant property damage	15. Commerce and Trade	15 U.S. Code §1193(a)
switchblade knife	15. Commerce and Trade	15 U.S. Code §1241(b)
unreasonably low prices	15. Commerce and Trade	15 U.S. Code §13a
small manufacturer	15. Commerce and Trade	15 U.S. Code §695 (e)
digital musical recording	17. Copyrights	17 U.S. Code §1001(1)
preexisting work	17. Copyrights	17 U.S. Code §101
audiovisual work	17. Copyrights	17 U.S. Code §102(a)
essential step	17. Copyrights	17 U.S. Code §117(a)(1)
technological measure	17. Copyrights	17 U.S. Code §1201(a)(3)(B)
familiar symbol	17. Copyrights	17 U.S. Code §1302
semiconductor chip product	17. Copyrights	17 U.S. Code §901(a)
independent economic value	18. Crimes and Criminal Procedure	18 U.S. Code §1839
fermented liquor	18. Crimes and Criminal Procedure	18 U.S. Code §1263
aural transfer	18. Crimes and Criminal Procedure	18 U.S. Code §2510(1)
leadership role in an organization	18. Crimes and Criminal Procedure	18 U.S. Code §5032
nonindustrial use	19. Customs Duties	19 U.S. Code §2802(7)
distributive share of the income	26. Internal Revenue Code	26 U.S. Code §1441(b)
nonmonetary benefits	28. Judiciary and Judicial Procedure	28 U.S. Code §1713
common business purpose	29. Labor	29 U.S. Code §203(1)(r)
dependent on hours worked	29. Labor	29 U.S. Code §207(e)
unduly disrupt the operations	29. Labor	29 U.S. Code §207(o)(5)
final average compensation	29. Labor	29 U.S. Code §623(i)(10)(iv)
standard coin	31. Money and Finance	31 U.S. Code §5151(b)
useful improvement	35. Patents	35 U.S. Code §101
basic allowance for subsistence	37. Pay and Allowances of the Uniformed Services	37 U.S. Code §101(25)
residential dwelling	42. The Public Health and Welfare	42 U.S. Code §5502(3)
mechanical recordation	44. Public Printing and Documents	44 U.S. Code §2201(1)
stored electronically	44. Public Printing and Documents	44 U.S. Code §4104
navigation equipment	46. Shipping	46 U.S. Code §70001(a)
substantial portion of the public	47. Telecommunications	47 U.S. Code §1401(8)
preemployment testing	49. Transportation	49 U.S. Code §31306(1)
gas pipeline facility	49. Transportation	49 U.S. Code §60101
hazardous liquid	49. Transportation	49 U.S. Code §60101

Table 3: The table lists all 42 terms included in the data set. The full citation to the source provision as well as the Title to which it belongs are provided.

all but one of the students took part.[7] At the meeting the students were first introduced to the background of the project, the end goal of my research, and the role their work would play in it. Then they were handed the printed version of the extended annotation guidelines (Appendix B) and the electronic version was sent to their emails. Finally, I walked them through the guidelines answering questions as they appeared. I also explained the logistics of the project, including the annotation environment (Google sheet), reporting, and the salary.

2. **Small training batch** – Following the initial meeting each student was assigned 100 sentences to label. For each student the sentences were related to a single term of interest. Each batch of sentences was assigned to exactly two students—these were related to the 'audiovisual work,' 'aural transfer,' 'preexisting work,' 'small manufacturer,' and 'technological measure.' The batch related to 'electronic signature' was assigned to the one student who did not participate in the initial meeting. The students were instructed to finish the labeling before the second meeting which means they had about 5 days for the work. They could perform the work at the time and place of their choosing. I emphasized that it was important they perform the work alone. In case of any questions they were supposed to turn to me. They were instructed to rely heavily on the annotation guidelines until they internalized them later.
3. **Second meeting** – The training continued with a 60 minutes long group meeting in which all of the students took part. At this meeting the students were paired with respect to the small training batches they were assigned (i.e., the students with the same batch were in the same pair). They were instructed to go through their labels together, discuss situations where they differed, and settle on a consensus label. They were also instructed to look for any systematic disagreement, i.e., repeated disagreements with the same underlying cause. My intention was to make students think about their annotation choices and ask them to explain these in front of their colleagues. At the end of the session each pair explained the systematic differences they identified to the whole group. The one student who did not participate in the initial meeting was paired with my advisor.
4. **First round of feedback** – The student annotators were provided with detailed feed-

[7] The student was supposed to replace the training with self-study.

back regarding their consensus labels produced at the second meeting (see above). The feedback was sent via email. It commented on any sentence I believed should have been labeled differently from what the students had agreed on. Each comment consisted of the suggested label and a list of reasons why I believed the suggested label to be more appropriate than the one selected by the students.

5. **Large training batch** – Following the second meeting each student was assigned 1,000 sentences to label. This time the students could be assigned with sentences related to several terms of interest. Each sentence was still assigned to exactly two students. The sentences were related to 'audiovisual work,' 'aural transfer,' 'electronic signature,' 'fermented liquor,' 'preexisting work,' 'small manufacturer,' 'switchblade knife,' and 'technological measure.' The students were instructed to finish the labeling before the final meeting which means they had about 2 weeks for the work (Spring holiday). As before, they could perform the work at the time and place of their choosing. Again, I emphasized that it was important they perform the work alone and in case of any questions they were supposed to turn to me. This time the students were instructed to focus on internalizing the annotation guidelines so that they needed to refer to them as seldomly as possible. The students were also asked to identify two sentences of each value type that they believed to be very clear examples of the respective value category as well as three sentences that they believed to be difficult to classify (i.e., 11 sentences in total).
6. **Final meeting** – The training continued with a 60 minutes long group meeting in which all but one of the students took part.[8] At this meeting each of the students presented the sentences he or she identified as either clear examples of a respective category or as difficult to classify. I provided feedback on each selection. The other students occasionally commented as well. The point of this exercise was to make sure that students were capable of recognizing sentences that are easy to classify and explain why is it the case. At the same time I wanted the students to understand why it might sometimes be difficult to decide on a category for a sentence.
7. **Second round of feedback** – The final component of the training was a brief feedback on the work the students did on the large training batches. I went through the annota-

[8]Note that this was a different student from the one who missed the initial meeting.

tions and looked for systematic problems, i.e., multiple mistakes that appeared to have a similar cause. The feedback was sent via email. It was mostly brief and in case of some students no problem was detected and therefore no feedback was necessary. Once this round of feedback was finished the students were considered fully trained.

4.5 ANNOTATION PROCESS AND INTER-ANNOTATER AGREEMENT

The first stage of the annotation process consisted of the training described above. Between February 28, 2019 and March 4, 2019 the students produced 1,166 annotations (around 100 per student) for 636 sentences (530 double annotated; 106 single). To measure inter-annotator agreement Krippendorff's α [61] was used. The measure is designed for situations of more than 2 annotators, any type of labels (including ordinal as in this case), and missing data. The α among the six pairs of students (names are anonymized with animal nicknames) for six different queries is reported in Figure 13 (one of the 'electronic signature' batches was annotated later). The overall α was 0.47. The agreement of each student to the consensus labels (α_c) produced by me and my advisor is reported in Figure 14.[9]

Low agreement between Emu and the consensus labels (α_c) is likely related to the student not attending the first meeting, i.e., attempting the task with no training at all. This also translates into the low agreement for the 'electronic signature' (Figure 13). The 'aural transfer' turned out to be a challenging term. The source provision was a definition of a "wire communication" in terms of the "aural transfer." What makes this complicated is that the statutory definition of aural transfer is quite close to the one of the "wire communication" (see Figure 15 for details). It appears the student annotators (Herring and Seal) often got confused and marked the definition of aural transfer ('high value') as having 'no value.' On top of that, Seal had trouble interpreting the rules for assigning the high value label in general. This resulted in the student annotating almost all the sentences with that label. These systematic errors translated into negative α_c for both students as well as low α.

Following the second meeting on March 4 and the first round of feedback, the students

[9]The process of creating the consensus labels is described later.

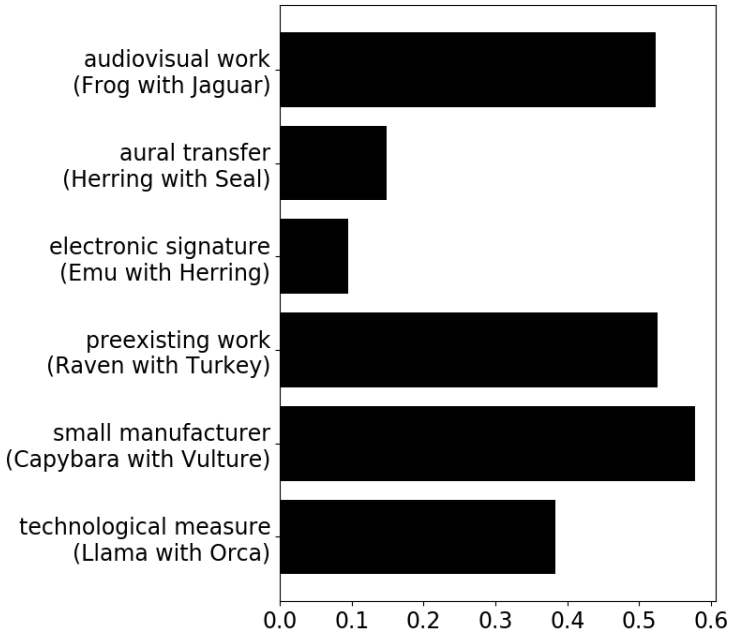

Figure 13: The figure shows inter-annotator agreement (α) between the six pairs of students on the annotations from the first training round.

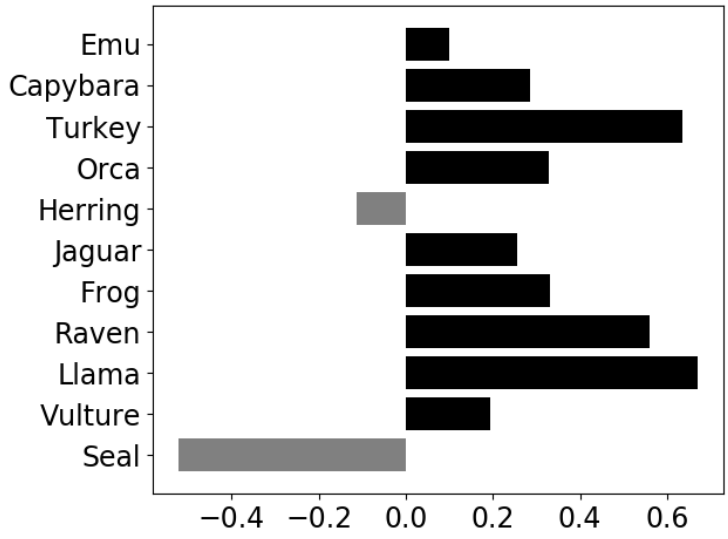

Figure 14: The figure shows the agreement (α) between the students and the consensus labels from the first training round.

"wire communication" means any aural transfer made in whole or in part through the use of facilities for the transmission of communications by the aid of wire, cable, or other like connection between the point of origin and the point of reception (including the use of such connection in a switching station) furnished or operated by any person engaged in providing or operating such facilities for the transmission of interstate or foreign communications or communications affecting interstate or foreign commerce;

"aural transfer" means a transfer **containing the human voice** at any point between and including the point of origin and the point of reception;

Figure 15: The figure shows that the statutory definition of "wire communication" is quite close to the definition of "aural transfer." This made the term challenging for the student annotators.

received over the email, each student was assigned with a second batch of sentences on March 6, 2019. Note that one of the students (Emu) dropped from the study after the first round. Before March 19, the students produced 11,564 annotations (approximately 1,000 per student) for 5,782 sentences (double annotated). This time each student received sentences related to 2–4 terms of interest (14 pairs altogether). The agreement among the 14 pairs of students for eight different queries is reported in Figure 16. The agreement to the consensus labels is reported in Figure 17.

The overall agreement of student pairs' labels and the final adjudication labels (see below) improved from 0.36 (first round) to 0.47 (second round). Interestingly, the agreement among the students themselves decreased from 0.47 to 0.33. These changes may reflect the fact that some students appear to have made a significant progress in understanding the task (Llama, Frog) while some continued to struggle (Capybara). For example, Capybara adopted an erroneous policy where sentences that cite case law or statutory law were almost always labeled as valuable (certain or high). This severely affected the student's α_c (Figure 17) as well as α on the 'electronic signature' and 'technological measure' batches. In general, the two main causes of a lower agreement for some of the queries are again the higher complexity of the query itself and/or a query being assigned to a lower performing student. Despite

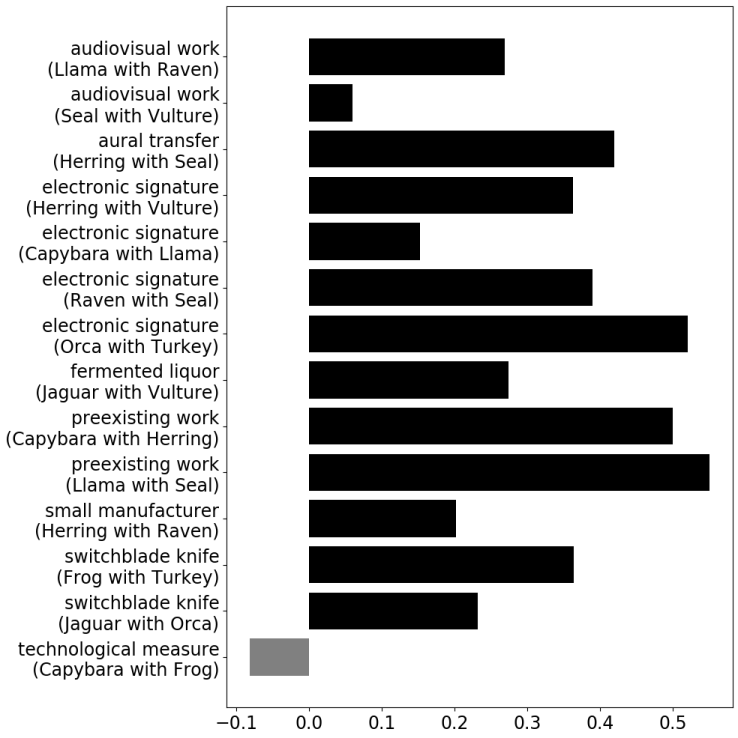

Figure 16: The figure shows inter-annotator agreement (α) between the fourteen pairs of students on the annotations from the second training round.

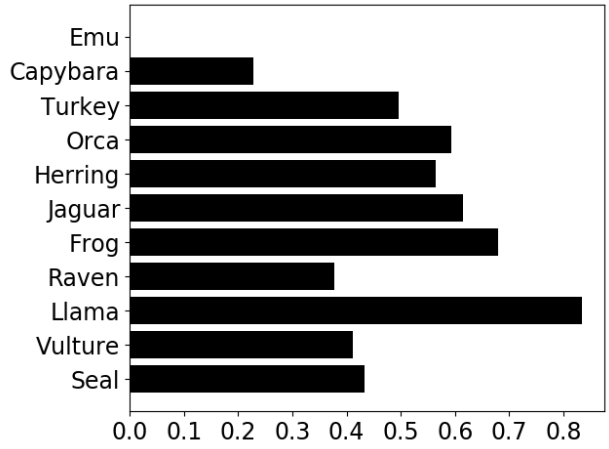

Figure 17: The figure shows the agreement (α) between the students and the consensus labels from the second training round.

the two issues being intertwined one can use Figures 16 and 17 to disentangle them to a certain degree. Take the example of Capybara and Raven—the two least performing student annotators. Both of them have low agreement with the final consensus labels ($\alpha_c < 0.4$ per Figure 17). It has already been explained that Capybara performed poorly. In addition, this is also suggested by the inter-annotator agreement scores from the batches assigned to the student as reported in Figure 16. In all three of Capybara's batches the α is somewhat lower than in other batches for the corresponding terms of interest. However, this is not the case for the batches assigned to Raven. Although, we cannot rule out that Raven is an under performer it certainly looks plausible that the reason for the low α_c is the student having been assigned with more difficult queries on average.

After the feedback received during the third meeting on March 19 as well as via the email, the students were assigned additional batches to work on. A batch typically consisted of about 1,000 sentences. After a student finished a batch he or she could ask for another one until the annotation budget was spent. Ideally, I would have liked to maximize:

- the number of queries each student worked on,
- the number of sentences annotated by high-performing students, as well as
- the number of different annotator pairs.

Realistically, I had to assign a batch to whichever student was free at a time. Since the study was conducted during the course of a semester, some students got busy and did not work on the task for a long time while others dropped almost immediately. Each batch could only contain sentences from a limited number of queries because I was not sure how many queries the students would be willing to annotate eventually.

Between March 19 and May 25 the students produced additional 31,918 annotations (16,012 sentences). While some of the students contributed large amount of work others dropped almost immediately. The overall agreement among the students for the whole data set including the annotations produced during the training is 0.31.[10] Table 4 summarizes the amount of work each student did. Except one, every student annotated at least 1,500 sentences. Some of the students managed over 6,000. Table 5 shows how many different pairs

[10]We investigate the causes of the low agreement and explain remedial steps below.

Annotator	# terms	# sentences
Capybara	26	6548
Emu	1	106
Frog	4	1765
Herring	17	6514
Jaguar	7	4349
Llama	16	6701
Orca	18	7041
Raven	10	4521
Seal	4	1583
Turkey	5	2313
Vulture	9	3201

Table 4: The table summarizes the amount of work produced by the individual student annotators.

of students I was able to create while navigating the contingencies of the whole annotation process. Overall I was able to cover 31 of 55 possible pairs. The minute details of the whole annotation process including all the α among the student pairs for each batch are reported in Appendix C.

4.6 ADJUDICATION

My original plan was to accept any label as correct where the students agreed. Where the students disagreed I or my adviser would have adjudicated the sentence and assigned a consensus label. However, in light of the low α the plan appeared problematic. In [104] the average inter-annotator agreement (weighted κ) between me and my adviser was at 0.66 (only three terms comprising 243 sentences). We noticed the lower agreement in the case

	Emu	Frog	Herring	Jaguar	Llama	Orca	Raven	Seal	Turkey	Vulture
Capybara	-	514	2780	508	348	1005	715	-	59	620
Emu	-	106	-	-	-	-	-	-	-	
Frog	-	113	-	-	-	-	-	463	675	
Herring	962	1477	640	344	139	-	367			
Jaguar	1046	468	428	-	-	825				
Llama	2074	434	411	1112	-					
Orca	2079	-	577	-						
Raven	319	102	-							
Seal	-	714								
Turkey	-									

Table 5: The table reports the existing student annotator pairs and the number of sentences their both annotated.

of "independent economic value" which was most likely due to the fact that the term was the first we were dealing with. Although, a detailed explanation of the annotation task was available, we did not practice before starting with the annotation. Based on this experience multiple measures to ensure high-quality of the annotations were adopted in [108]. Here, the annotators were I, my adviser, and the third co-author (a graduate student with a completed law degree). The final labels were either assigned by consensus of all three of us or on the basis of the third annotator independently resolving differences between the other two annotators' sentence labels. For each term we selected a subset of sentences that were labeled first. Once these subsets were finished, the annotators met to resolve any systematic disagreements. Only after that meeting did we proceed to label the rest of the sentences. We confirmed that the task is challenging and requires substantial annotator training. After the first labeling round on "independent economic value" the α was only 0.30. After the resolution of the systematic differences, however, it rose to 0.55. Because of the strategy for resolving systematic disagreements early in the annotation process, we achieved higher

overall agreement ($\alpha = .79$) than in [104].

The three annotators in [108] received a significant amount of training. My adviser and I designed the task and participated in the previous annotation round for [104]. Although I did not keep time records, it seems plausible that each of us spent tens of hours on activities that could be considered training. Similarly, the third annotator participated in full-afternoon long training meetings that amounted to about 10 hours in aggregate. Given the available amount of resources and other practical constraints, it was not possible to provide the 11 student annotators with the same level of training. As described in Section 4.4 most of the students participated in 3 meetings that amounted to 3 hours in total. My expectation was that while the agreement was bound to be somewhat lower than in [104] and [108], it would have most likely fallen to $0.5 < \alpha < 0.6$.

The actual agreement of $\alpha = 0.31$ had several implications. First, the number of sentences where the student annotators disagreed and which required adjudication was rather high. In total 7,262 sentences required adjudication (44%). Another issue was the potentially problematic quality of the annotations where students did agree. Additionally, the low α gave rise to important questions about the task.

In order to ensure the high quality of the resulting labels and to investigate the causes of the low α, I decided to abandon the original plan of accepting the labels where the students agreed. This meant a significant increase in the workload for the two adjudicators (me and my adviser) where we had to go over all the 15,906 sentences. An equal distribution of the task translated into approximately 80 hours of adjudication work for each of us. However, given the situation this appeared to be necessary. We carried out the work in June and July 2019. Apart from obtaining the final consensus labels for all the sentences, such an adjudication process also enabled us to investigate the causes of the low α. Note that an alternative approach could have been to accept that there might be problematic labels, presumably supplied by problematic annotators. Then, instead of dealing with the task in terms of having so-called gold labels, I could have employed strategies to deal with data coming from multiple annotators such as in [85] or [114]. However, as this is quite a complex issue, I did not want to divert my attention from the main objective of the work and hence, did not follow this path.

The adjudication process was carried out in three stages:

1. *Disagreement Resolution* – In this stage we dealt with the 7,262 sentences where the students disagreed. The terms of interest were divided among the two of us so that the amount of work was roughly equal. Note that we did not constrain ourselves to select one of the labels proposed by the students. This means that the adjudicator could opt for a label none of the students had selected.

2. *Agreement Quality Check* – In this stage the remaining sentences, i.e., the 8,644 where the students agreed, were checked by the two of us. Importantly, each of us was assigned with those terms of interest the other one adjudicated in the first adjudication round. These also contained the consensus labels from the first round. The task was not only to check the sentences where the students agreed but to also spot-check the consensus labels produced by the other adjudicator. The goal was to ascertain that no systematic discrepancies between the two adjudicators existed. No such discrepancy was detected.

3. *Consistency Check* – We applied the simhash algorithm from [72] to group identical and nearly identical sentences into near duplicate clusters.[11] For all the clusters with size greater than three we checked if all the sentences had the same labels. In case of a difference we either made sure that the difference was justified or corrected the mistake. By applying this approach we were able to correct a number of obvious mistakes due to fatigue/lack of focus and harmonize edge cases.

After the above three steps were performed, the 15,906 sentences resulting from the non-training part of the student annotations were assigned their final consensus labels. For completeness, it is important to mention that there were also 6,418 sentences involved in the first and second training round. These were adjudicated during the training in order to provide students with feedback. Finally, there were the 4,635 sentences from [108]. These additional 11,053 sentences did not go through the first two steps of the above described adjudication process. However, they went through the third consistency check step. In this way the consensus labels for all the 26,959 sentences were finalized.

[11] We used the implementation of the algorithm from github.com/leonsim/simhash.

4.7 ANNOTATION ANALYSIS

The importance of analyzing the shortcomings in an annotation process is emphasized in [4]. There are three possible causes that could have led to the low α:

1. Some of the annotators were not sufficiently trained to perform the challenging task.
2. The task is not well-defined by the annotation guidelines. Some aspects of the task are defined vaguely, leaving too much space for a subjective judgement of an annotator.
3. The task itself is not very objective. The analyzed phenomenon is quite subjective and different people may legitimately hold differing opinions as to the value of each sentence.

Appendix C reports each student's agreement with the final consensus labels α_c for each individual batch. Several terms of interest stand out because α_c of all the students participating in the annotation of the terms suggest that there is no agreement or even systematic disagreement. The terms as well as assumed causes of the lack of agreement are:

- *aural transfer* – This query has been already analyzed. As shown above it is very challenging. The students were clearly not sufficiently trained to tackle this term in the first training round as they were asked to.
- *essential step* – A large number of sentences mention the term in a completely different meaning. Therefore, they should be labeled as having 'no value.' This is almost universally not recognized by the students resulting in a strong systematic disagreements with the final consensus labels. This phenomenon is problematic because it manifests only in a small number of terms. It was not encountered much during training. Therefore, students were not well prepared to tackle it.
- *significant property damage* – This term appears to be poorly chosen. Most of the sentences appear to have little value (i.e., 'potential value' label). The extreme scarcity of the more valuable sentences means that if the students missed some of them the agreement was very low. That is exactly what happened here.
- *substantial portion of the public* – This query was challenging because many sentences used the term in a different meaning. One of the students recognized the phenomenon and judged the meaning as completely different (hence a lot of 'no value' labels). The

other student missed the phenomenon. The adjudicator mostly understood the meaning as different yet related (hence a lot of 'potential value' labels).

The above analysis shows that due to a number of reasons some of the terms were very challenging. This could be due to a more complicated term being tackled in an early stage of the annotators' training or due to reasons that made the term complicated even for the annotators that were already trained. This suggests deficiencies in the training process as well as the assignment of terms to annotators that were not yet prepared to tackle them. In addition, the analysis uncovered a more general problem—sentences that contained mentions of the respective term with a different meaning. Here, it appears the issue may be more serious than somewhat inadequate training. The decision about the appropriate label for a sentence is driven by the rules described in subsections B.3.2 and B.3.3 in Appendix B. On a second look it appears that I left these rules quite vague. Consider the following excerpts:

> "[...] There are ways to argue that [the terms] mean the same thing, but also ways to argue that [they] mean something different. The decision in such cases is left to the annotator's discretion. [...]"

> "[...] The decision as to the degree of the difference in such cases is left to the annotator's discretion. [...]"

Retrospectively, it appears that providing such vague definitions was a mistake. These should have been specified in terms of more definite rules. However, it may also be the case that an objective assessment of this phenomenon is too difficult if not impossible. Therefore, it might have been more prudent to approximate it through other means. Perhaps, I could have considered a term to have different meaning in any case where it would appear the term does not come from the source provision (e.g., a different area of law would automatically trigger the rule).

Above we have discussed the terms of interest that appear to be problematic in general, i.e. the ones for which all the students had low α_c. A related phenomenon is introduced by the terms where only some (but not all) students struggled. These include the following:

- *dischargeable consumer debt* – Vulture systematically classified a larger number of sentences that were verbatim quotations of the part of the source provision as having 'certain

value.' Yet these are very clear instances of sentences with 'no value.' This was correctly recognized by Capybara.

- *electronic signature* – Emu attempted to annotate the first training batch while not attending the first training session. This was already discussed above.
- *hybrid instrument* – This appears to be quite a challenging term. It comes from a specialized financial domain (commodity exchanges) and the sentences contain a lot of technical jargon. Capybara classified a number of sentences as 'high value' although they clearly did not meet the definition.
- *nonindustrial use* – Here the issue was the relatedness of the meaning between the term of interest from the source provision and its mentions in the sentences. Both students correctly recognized that the meaning was different in case of many sentences. However, Herring chiefly deemed the meanings as related while Capybara as well as the adjudicator thought that the meanings were completely different.
- *security vulnerability* – Orca failed to recognize most of the 'high value' sentences, annotating almost all of them as having 'potential value.' This appears as a possible instance of not paying enough attention to the task (fatigue?).
- *stored electronically* – Here the issue is identical with the one described above. Orca failed to recognize most of the 'high value' sentences. Interestingly the student worked on these two batches at a very similar time.
- *technological measure* – This is a batch that was assigned during the second round of training. Capybara adopted an erroneous policy of marking any sentence that quoted an official authority as having 'certain' or 'high value.' This led to a systematic disagreement with the other student annotator as well as with the adjudicator. This issue was already described above.
- *unduly disrupt the operations* – There are many instances of a verbatim quotation of the source provision with additional short sequence of words added. Because there is the additional information these sentences should be labeled as having 'potential value.' Capybara consistently annotates these as having 'no value' (likely due to a lack of focus).
- *unreasonably low prices* – Here Capybara failed to notice that in many sentences the term of interest appears in a very different meaning. Therefore, he or she consistently

rated the 'no value' sentences much higher.

- *useful improvement* – This term was quite challenging. There were many sentences following a similar pattern that were mostly deemed to have 'certain value' by Capybara, Jaguar, Orca, and Turkey, as well as the adjudicator. However, Herring and Llama annotated these sentences as having 'potential value.' Because the number of these sentences is quite high, the α_c scores of Herring and Llama are very low.

The terms that are problematic for just some of the students are again manifesting the different/related meaning problem identified earlier. This phenomenon definitely points to inadequacies of the training provided to the students. It almost certainly also points to a flaw in the annotation guidelines. Finally, we cannot rule out the possibility that this phenomenon cannot be assessed objectively which would point to a flaw in the task definition itself. I have identified a couple of instances where students simply missed an obvious phenomenon pertaining to a larger number of sentences which resulted in a very poor α_c scores. Some of the terms were simply quite complicated.

Thirdly, I noticed that some of the students produced higher quality annotations than others, i.e. annotations that had higher agreement with those produced by the adjudicators. The raw α_c scores are reported in Figure 18. It appears that Llama, Frog, and Turkey did much better than Emu, Orca, and Herring. This is important because it suggests that some students might have required additional attention during training. If such attention would have been provided their performance might have been closer to the top performing annotators. Consequently, the agreement might have been higher. The raw scores are somewhat problematic because they disregard the fact that some of the students might simply be assigned with more challenging terms on average. In Figure 19 I report weighted average difference to the mean α_c (per term). A positive difference suggests that on average a student did better than his peers. Although, still not perfect this statistic takes the overall difficulty of a term into account. For example, Lamma has $\alpha_c = 0.64$ on "semiconductor chip product" while Herring has $\alpha_c = 0.89$. At the same time, Raven has $\alpha_c = 0.50$ on "small manufacturer" topping the other three annotators who have $\alpha_c < 0.40$. Using the difference we would estimate that Herring and Raven are the best annotators taking the assumption that "semiconductor chip product" was a less challenging term than "small manufacturer."

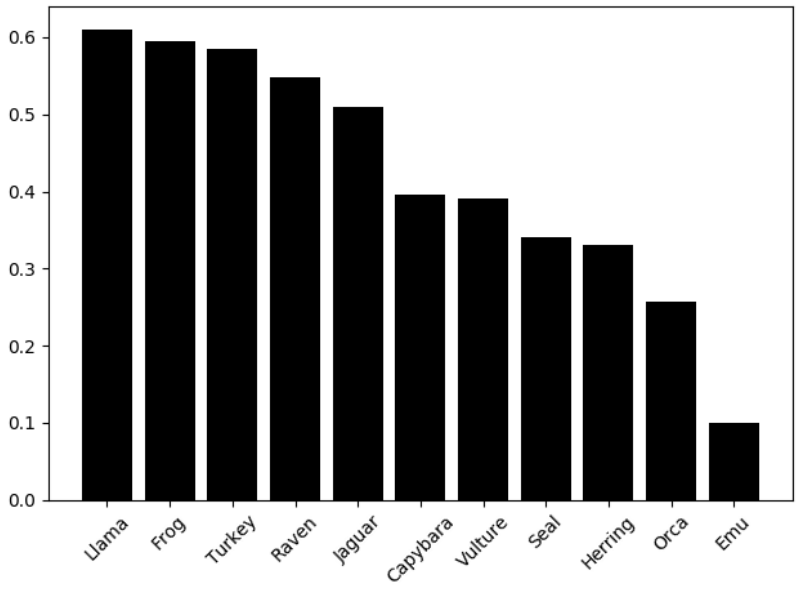

Figure 18: The figure shows the average agreement between the students and the final consensus labels. It is clear that some students performed much better than others.

Figure 19 largely confirms what is shown in Figure 18. It appears that Frog might have been the best trained annotator overall by quite some margin. Orca appears as a rather average performing annotator as opposed to one of the worst as suggested by Figure 18. Most importantly, Figure 19 confirms that the differences in quality among the student annotators do exist.

In conclusion, the 11 law student annotators produced over 40,000 annotations for more than 20,000 double-annotated sentences over the period of three months. The overall inter-annotator agreement ($\alpha = 0.31$) was rather low. In order to understand the causes of the low α and to ensure high quality of the resulting labels, each sentence was re-evaluated by me or my adviser. The resulting consensus labels were then instrumental in understanding why the α was low. I have potentially identified one serious issue with the task definition. The issue is related to the assessment of the difference and relatedness of the meanings of an identical term used in different contexts. It is clear that at minimum the instructions provided in the annotation guidelines regarding this phenomenon are too vague and they would require

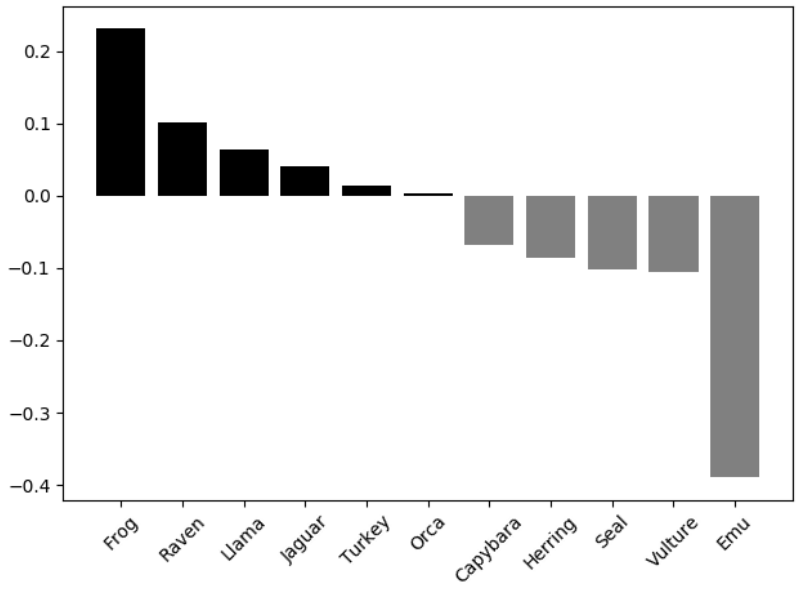

Figure 19: The figure shows the difference between the student agreement to the consensus labels and average agreement per query. Positive values indicate that student performed better than all the students on average.

further specification. It is equally clear that the students did not receive adequate training to tackle this phenomenon. This is mostly due to the fact that the phenomenon did not manifest itself on the terms selected for training. In addition, I have established that some of the students performed much better than others. It appears reasonable to assume that the low agreement might be partially due to inadequate performance of some students, which might have been corrected by additional training.

4.8 RESULTING DATA SET

The annotation effort described in Sections 4.1–4.7 resulted in the data set that consists of 26,959 sentences. These are annotated in terms of the four categories with respect to their usefulness for the interpretation of the corresponding statutory term (‘high value,’ ‘certain

value,' 'potential value,' and 'no value'). The sentences are responses to the 42 queries (39 new and the three original) and they come from 20 different areas of law (Titles of the U.S.C.; see Table 3). As mentioned before, the documents were indexed at the level of full cases as well as segmented into the head matter and individual opinions (e.g., majority opinion, dissent, concurrence). I used the sentence segmenter from [107] to segment each case into individual sentences and indexed those as well. As before, I used the sentences to create paragraphs. The detailed statistics on the counts of these segments per query are reported in Table 6.

From the statistics reported in Table 6 it is clear that the subsets corresponding to the individual query terms are quite diverse. Some of the terms have several thousand sentences that come from thousands of cases while responses to other terms comprise only a handful of sentences. For example, on the one side there are sentences related to "useful improvement" (3,867) or "residential dwelling" (3,235) and sentences related to the "mechanical recordation" (18) or "semiconductor chip product" (25) on the other side. It is important to emphasize that the terms were not chosen randomly. As explained in Section 4.3 I could only pick terms that did not return too many results. This led me to accept only terms that returned less than 5,000 sentences.

I balanced between an interest in the inclusion of some terms with a larger number of related sentences and an interest of having as many query terms as possible. On the one hand, using only terms with a small number of sentences, say up to 50, I could have ended up with a data set that would comprise over 500 queries. Although, this would have been much nicer than the actual 42, the data set would have a serious flaw. Certain phenomena (discussed below) manifest only for terms that have many related sentences. A data set consisting of only queries returning a small number of sentences would completely miss these. The main point is that the resulting data set has no claims of being a true representation of the "real query term–mentioning sentences" distribution. It is simply my best attempt to have a data set with a larger number of queries while some of them have many (thousands) of related sentences. Due to the cost of labeling I could not include even a single term with tens of thousands (or more) of sentences.

Beside the raw numbers of cases, opinions, paragraphs, and sentences I also report the

	cases	opinions	paragraphs	sentences	s/c ratio
Accommodation Trade	26	26	60	69	2.65
Audiovisual Work	515	520	1059	1261	2.45
Aural Transfer	88	92	130	139	1.58
Basic Allowance For Subsistence	54	54	73	79	1.46
Common Business Purpose	371	377	735	880	2.37
Cybercrime	51	52	63	71	1.39
Dependent On Hours Worked	13	13	23	27	2.08
Digital Musical Recording	8	8	36	43	5.38
Distributive Share Of The Income	143	146	166	172	1.20
Dischargeable Consumer Debt	119	119	133	135	1.13
Electronic Signature	575	580	1241	1581	2.75
Essential Step	2026	2071	2307	2374	1.17
Familiar Symbol	55	55	63	64	1.16
Fermented Liquor	1158	1320	1907	2133	1.84
Final Average Compensation	93	98	176	210	2.26
Fully Amortize	285	290	401	421	1.48
Gas Pipeline Facility	48	48	60	66	1.38
Hazardous Liquid	162	164	310	359	2.22
Hybrid Instrument	47	48	81	87	1.85
Identifying Particular	1882	1892	2134	2217	1.18
Independent Economic Value	872	888	1435	1538	1.76
Leadership Role In An Organization	26	26	30	30	1.15
Mechanical Recordation	14	14	16	18	1.29
Navigation Equipment	118	119	150	154	1.31
Nonindustrial Use	22	22	28	32	1.45
Nonmonetary Benefits	127	127	185	204	1.61
Preemployment Testing	43	46	66	70	1.63
Preexisting Work	401	403	712	823	2.05
Residential Dwelling	2280	2298	3030	3235	1.42
Security Vulnerability	56	57	82	84	1.50
Semiconductor Chip Product	10	10	23	25	2.50
Significant Property Damage	207	210	221	223	1.08
Small Manufacturer	354	360	427	452	1.28
Standard Coin	91	99	161	179	1.97
Stored Electronically	191	192	228	232	1.21
Substantial Portion Of The Public	227	231	337	366	1.61
Switchblade Knife	860	888	1464	1646	1.91
Technological Measure	101	104	464	616	6.10
Unduly Disrupt The Operations	28	30	46	48	1.71
Unreasonably Low Prices	334	340	485	508	1.52
Useful Improvement	2972	3167	3736	3867	1.30
Viticultural	95	98	196	221	2.33
Total	17148	17702	24680	26959	1.57

Table 6: The table reports how many cases, opinions, paragraphs, and sentences are related to each of the 42 queries. The s/c ratio is an average number of sentences mentioning the term of interest per case.

sentence to case ratio in Table 6. This measure captures the average number of sentences mentioning the term of interest appearing in a single case. The ratio close to 1.0 suggests that most of the cases contain just one (perhaps incidental) mention of the term and there are only a few cases that contain multiple mentions. A higher ratio then means there are a lot of cases that mention the term multiple (perhaps many) times. Most of the terms have the ratio between 1.0 and 2.0. The smallest ratio of 1.08 belongs to "significant property damage." Here, almost all the cases contain just one sentence mentioning the term. Only a very small number of cases has two or three sentences. The greatest ratio of 6.10 is reported for the "technological measure." In case of this query almost all the cases have multiple sentences. Some of the cases contain as many as 25 sentences mentioning the term. Since the number of times a term is mentioned in a document is an approximation of how much the document is about the term, one could expect that the task of interpretive sentence retrieval might face quite a different set of challenges on the terms with a high sentence to case ratio as opposed to the terms where the ratio is low.

Figure 20 shows the distribution of labels for each of the terms. The Figure confirms the main characteristic apparent from the earlier iterations of the corpus ([104] and [108]) that the less valuable categories ('no value' and 'potential value' are dominant). For all the larger terms and almost all the small terms it holds that either the 'no value' or the 'potential value' category is the most numerous one. No matter the size, it is still the case that some of the terms contain quite a considerable number of more valuable sentences (e.g., "audiovisual work" or "switchblade knife") while others are significantly more limited in this respect (e.g., "essential step" or "hazardous liquid").

I noticed an interesting phenomenon that manifests especially clearly on the larger query terms. Consider the example of "essential step" and "residential dwelling." While both of the terms are among the most sizeable ones they are quite different in terms of the dominant sentence category. While "residential dwelling" is heavily dominated by the 'potential value' sentences the majority label for "essential step" is 'no value.' This is due to the fact that "residential dwelling" is a technical term that has a single unified or closely related meaning irrespective of the context. On the other hand, "essential step" is quite a general term that appears in many different contexts. For example, the term of interest was referring to

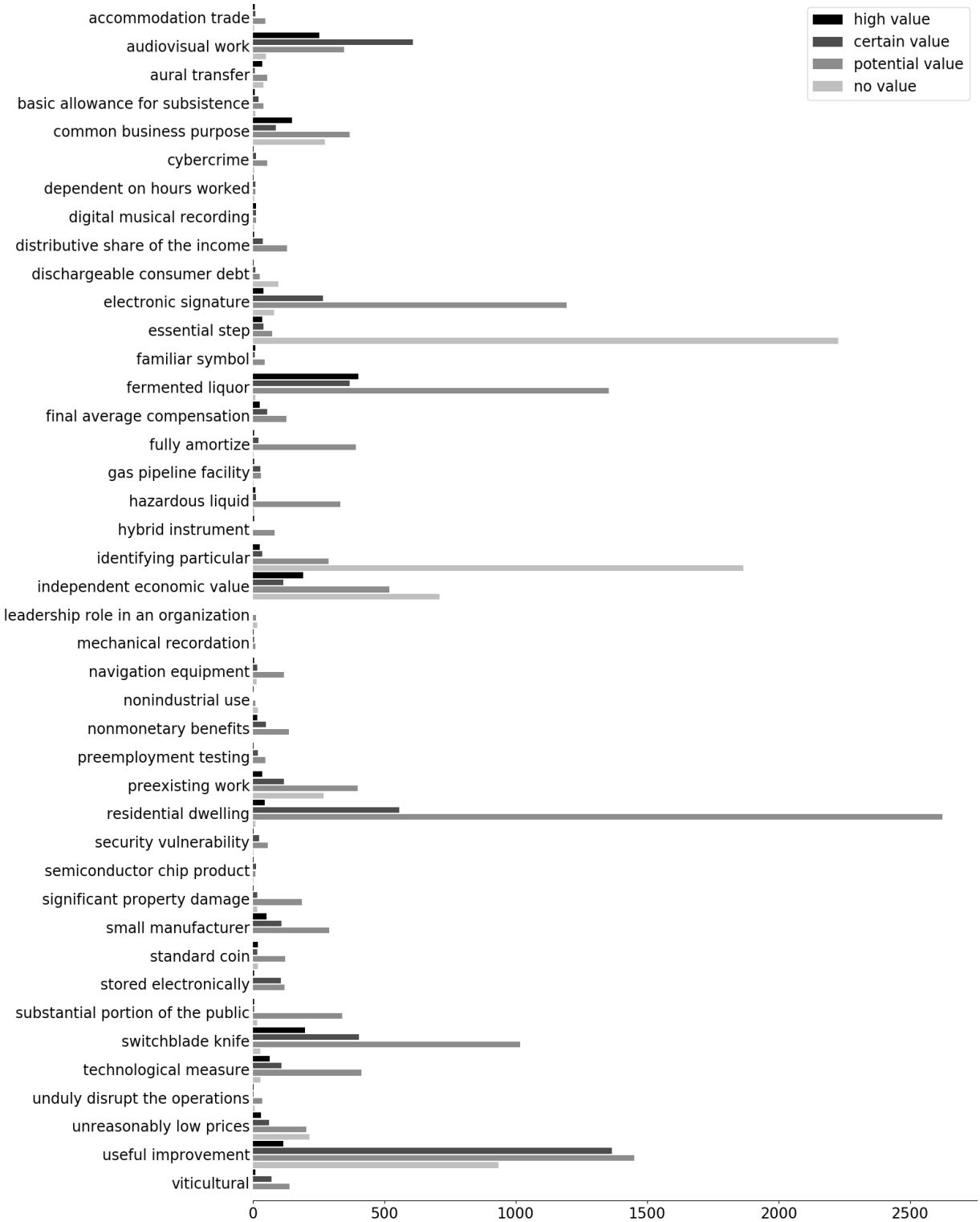

Figure 20: The figure shows the distribution of the labels for each of the 42 queries.

the "essential step" in the utilization of a computer program that would permit creation of a program's copy even without an explicit permission. The sentences often mentioned "essential step" taken in the course of court proceedings. The different contexts make the meanings completely unrelated. "Useful improvement" is somewhere in between with respect to this phenomenon.

5.0 EVALUATION METHOD

5.1 DIVISION OF THE DATA SET INTO FOLDS

In [38] it is explained that the so-called holdout method of separating the available data into a training set and a (disjoint) test set yields quite unreliable outcomes in IR. It is the case because the results may be heavily dependent on the split (e.g. the study in [96] shows that the performance on the two halves of the collection differed by 20%, and, for other splits, system differences sometimes were significant, and sometimes not). Therefore, cross-validation is recommended (see e.g. [124, p. 152–6]). Because in this work the number of queries is small the leave one out method would be optimal. In that case each single experiment would have to repeat the optimization/training 42 times which would be prohibitively time consuming. Therefore, I opted for a limited number of folds.

For the purposes of the experiments described in the following chapters the data set would need to be divided into k folds (subsets). Since there are huge differences among the individual queries and since the queries are not too numerous I needed to ensure that queries of different types make it to the individual folds. Essentially, what I would have liked to prevent is to, for example, allocate all the queries with a large number of sentences to one subset and all the smaller queries to the others. For this reason I used stratified sampling. Potentially there are many dimensions along which the queries could be assessed. Two very important ones are the size of the query (i.e., the number of sentences returned in a response to the query) and its richness.

Richness is a term often used in technology assisted review in e-Discovery referring to the prevalence of responsive documents in a collection. I adapt the idea for the circumstances of this work by defining a measure that describes the prevalence of valuable sentences in the

data set. First, I assign value to a sentence of each type on the scale from 0 to 10:

$$val(s_i) = \begin{cases} 10 & \text{if } s_i \text{ has 'high value'} \\ 5 & \text{if } s_i \text{ has 'certain value'} \\ 1 & \text{if } s_i \text{ has 'potential value'} \\ 0 & \text{if } s_i \text{ has 'no value'} \end{cases}$$

'High value' sentences get the highest valuation while 'no value' sentences the lowest. The reason why I use the scale of 0 to 10 and assign the value in the non-linear fashion is to overcome the dominance of the less valuable sentences. If I would simply use the scale from 0 to 3 the differences in richness would likely only reflect a shifting balance between the 'no value' and 'potential value' sentences. The contribution of the 'high value' and 'certain value' sentences would likely be minimal. Hence I assign the 'high value' and 'certain value' sentences with much more value than the 'potential value' and 'no value' sentences. It is important to emphasize that the scores do not really reflect the value ratio among sentences of different types. This means I am not implying that 'high value' sentences are 2 times more valuable than 'certain value' sentences or 10 times more valuable than 'potential value' sentences. In order to assess richness (R) I then simply compute an average value of a sentence within a query (q):

$$R(q) = \frac{1}{n} \sum_{i=1}^{n} val(s_i)$$

Figure 21 shows the terms of interest plotted along the two important dimensions (i.e. the size of a query and its richness). Because of the relatively small number of queries I cannot insist on too granular stratification. Therefore, I slice the space just once along each of the dimensions. There are two relatively compact clusters in the bottom part of Figure 21. On the richness dimension one can separate these by placing the border at $R(q) = 2.0$. For the size dimension I set the threshold at 550 which appears to be a reasonable upper bound for the two clusters. The queries are then divided into "small" ($size < 550$) and "large" ($size \geq 550$) as well as "sparse" ($R(q) < 2.0$) and dense ($R(q) \geq 2.5$). The dividing lines are not to be understood as an objective distinction between the categories. They simply

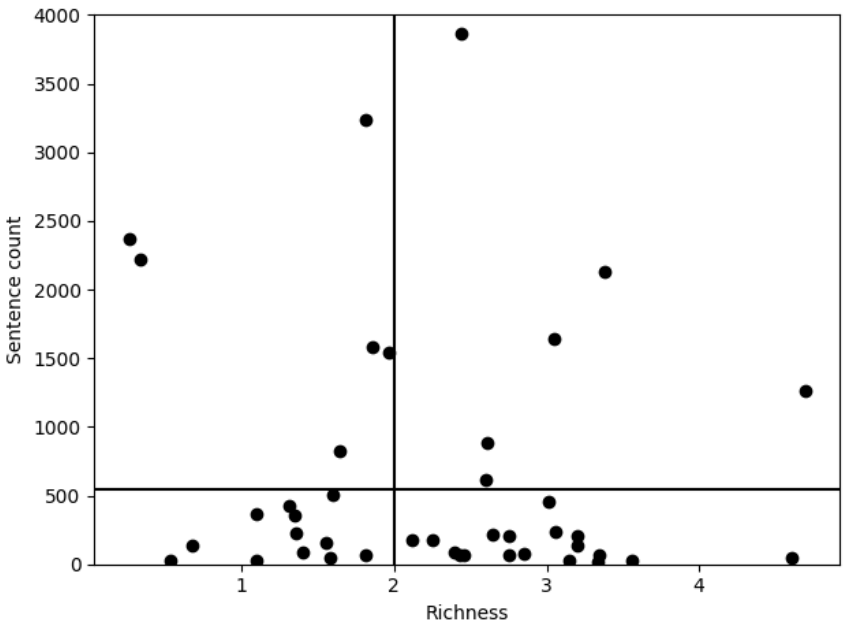

Figure 21: The figure shows the terms of interest plotted along the the size of a query (i.e., the number of retrieved sentences) and its richness.

appear to provide solid grounds for making sure I get balanced division into folds. Figure 21 then presents the stratification which divides the space into four categories:

- *Small and sparse queries* – This category corresponds to the lower left region of Figure 21. 12 terms belong in this category (28.6%). For almost all the terms in this category it holds that either the 'no value' or the 'potential value' sentences are completely dominant leaving very little space for sentences of other value classes. The more valuable sentences are rare and because of the limited size of these queries they would typically aggregate to single or lower double digit counts.
- *Small and dense queries* – The terms from this category lie in the lower right region of Figure 21. It is the most numerous category comprising of 18 terms (42.8%). The distribution of sentences into the value classes is more uniform for the terms in this category. Though, the sentences from the lower value classes are still numerous the dominance is not as strong as is with the terms in the first category described above.

electronic signature essential step identifying particular independent economic value residential dwelling preexisting work	audiovisual work common business purpose fermented liquor switchblade knife technological measure useful improvement
cybercrime dischargeable consumer debt fully amortize hazardous liquid hybrid instrument leadership role in an organization navigation equipment nonindustrial use significant property damage substantial portion of the public unduly disrupt the operations unreasonably low prices	accommodation trade aural transfer basic allowance for subsistence dependent on hours worked digital musical recording distributive share of the income familiar symbol final average compensation gas pipeline facility mechanical recordation nonmonetary benefits preemployment testing security vulnerability small manufacturer semiconductor chip product standard coin stored electronically viticultural

Table 7: The table shows the distribution of the terms into the four categories, i.e., small sparse (lower left), large sparse (upper left), small dense (lower right), and large dense (upper right).

- *Large and sparse queries* – This category corresponds to the upper left region of Figure 21. Out of the total 42 terms six falls into this category (14.3%). This category is similar to its "small counterpart" in the sense that either the 'no value' or the 'potential value' sentences are dominating the count. However, the more valuable sentences amount to a significant number because of the size of the queries (higher double or lower triple digits).

- *Large and dense queries* – This category corresponds to the upper right region of Figure 21. Six terms belong in this category (14.3%). This category is also similar to its "small counterpart" in the sense that the distribution of sentences into the four value classes is more uniform. Since the queries in this category have many retrieved sentences the total numbers of more valuable sentences are high (often hundreds).

Table 7 shows the distribution of the terms into the four categories. The placement of the regions corresponds to the placement in Figure 21. The typical division into 10 folds would be possible. However, a division into six folds appears more appropriate in this specific situation. This is because each of the above described query types is associated with a number of queries that can be nicely divided by six. This leaves me with folds where each includes:

- 2 small sparse queries,
- 3 small dense queries,
- 1 large sparse query, and
- 1 large dense query.

Hence, in my experiments I will use the division into six non-overlapping folds. In each iteration five folds will be used for training and validation (if needed) and one for testing. Note that for the experiments in Chapter 6 I will use only four of the folds, setting the second and sixth folds aside. This is useful because the main goal of Chapter 6 is to analyze the task for the purposes of understanding its fundamental nature as well as putting together the set of predictive features. Setting the two folds aside limits my ability to detect statistical significance (Section 5.3) in the Chapter 6 experiments. On the other hand, it enables me to asses how well do the methods built upon insights gained from those experiments generalize to unseen data. Specifically, I can analyze if the performance on the unseen data is much lower than the performance on the data the experiments were performed on. If this would be the case then it would suggest that the insights pertain to the specifics of the data set but may not be valid in general.

Table 8 reports the results of a random allocation of the terms into the individual folds. Figure 22 shows the allocation from the perspective of the sentences' values as well as from the perspective of the contribution of the individual types of queries. Despite the use of stratified sampling it is obvious that differences among the folds exist. For example, the third and the fourth folds appear to be clearly dominated by one sentence class whereas the other folds are distributed slightly more uniformly. Additionally, the first two folds are obviously smaller than the other ones. Also, the dominant sentence label in the last two

Fold 1	**Fold 2**	**Fold 3**
navigation equipment	nonindustrial use	unduly disrupt the operations
leadership role in an organization	significant property damage	substantial portion of the public
aural transfer	nonmonetary benefits	small manufacturer
semiconductor chip product	basic allowance for subsistence	accommodation trade
distributive share of the income	stored electronically	standard coin
preexisting work	independent economic value	residential dwelling
audiovisual work	technological measure	common business purpose
Fold 4	**Fold 5**	**Fold 6**
hazardous liquid	hybrid instrument	dischargeable consumer debt
fully amortize	unreasonably low prices	cybercrime
security vulnerability	gas pipeline facility	digital musical recording
familiar symbol	preemployment testing	viticultural
mechanical recordation	final average compensation	dependent on hours worked
electronic signature	identifying particular	essential step
fermented liquor	useful improvement	switchblade knife

Table 8: The table reports the results of a random allocation of the terms into the individual folds.

folds is 'no value' while the 'potential value' label is the most prevalent one in the remaining folds. These variations are mostly caused by large queries that may differ quite considerably as shown in Figure 21. For example, the dominance of the 'potential value' class as shown in Figure 22 in the third and fourth folds as shown in Figure 22 is due to the inclusion of "residential dwelling," "fermented liquor," and "electronic signature" that are all large and heavily dominated by sentences of this class. The smaller size of the first two folds is clearly due to inclusion of large queries that are smaller as compared to other large queries ("preexisting work," "audiovisual work," "independent economic value," and "technological measure"). The cause of the 'no value' sentences being the majority class in the last two folds are the "identifying particular" (fold 5) and "essential step" (fold 6).

The above shown differences are the consequence of a limited number of larger queries. This is something I have to accept due to a very high cost of labeling of large numbers of sentences related to these terms. At the same time, the stratification did really help to make the folds similar to each other in many important aspects. Despite the size differences exist these are not huge. The two smaller folds are each about half of the size of the bigger ones. Since I know the number of queries in each fold is the same the difference appears

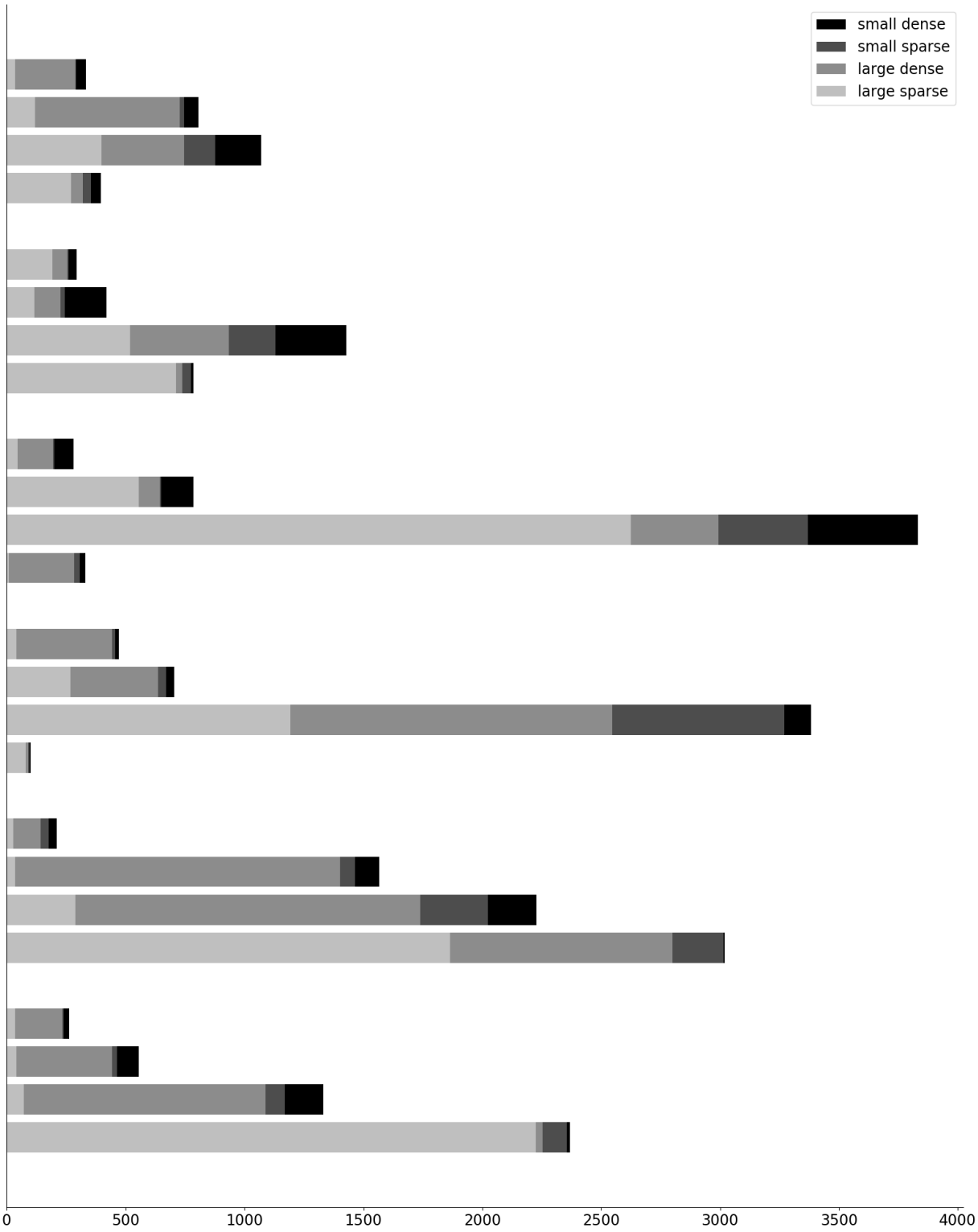

Figure 22: The figure shows the distribution of sentence value types into the individual folds. Contribution of different query types is shown as well.

to be acceptable. When it comes to a majority class in each term this is always one of the less valuable categories, i.e., either the 'no value' or the 'potential value' label. The 'high value' label is always the rarest. Given the circumstances, the stratification appeared to have worked fairly well producing reasonably balanced folds.

5.2 EVALUATION MEASURES

Since the notion of relevance in this work is non-binary I use normalized discounted cumulative gain ($NDCG$) to evaluate the performance of different approaches. An output of the presented ranking algorithms for each query q_j has the form of an ordered tuple of sentences $S_j = (s_1, s_2, \ldots, s_n)$. I chose to evaluate the rankings at $k = 10$ and 100 which means that the tuples produced by the algorithms are truncated to the respective lengths. Note that the chosen values of k are higher than typical. Measuring at $k = 100$ may even appear somewhat extreme. However, legal search differs from the general web search. Assuming a lawyer has confidence in the query (based on seeing several relevant hits towards the top of the results' list), he or she might be inclined to inspect the results way beyond what would a typical web search user do. For each query q_j the $NDGC$ at each k is then computed as:

$$NDGC(S_j, k) = \frac{1}{Z_{jk}} \sum_{i=1}^{k} \frac{rel(s_i)}{log_2(i+1)}$$

The function $rel(s_i)$ takes a sentence as an input and outputs its value in a numerical form. It is defined as follows:

$$rel(s_i) = \begin{cases} 3 & \text{if } s_i \text{ has 'high value'} \\ 2 & \text{if } s_i \text{ has 'certain value'} \\ 1 & \text{if } s_i \text{ has 'potential value'} \\ 0 & \text{if } s_i \text{ has 'no value'} \end{cases}$$

Z_{jk} is a normalizing quantity which is equal to $NDGC(S_j, k)$ where S_j is the ideal ranking. In case of this work this means that all the s_i with 'high value' labels are at the beginning

positions of the tuple, followed by those with 'certain value,' then 'potential value,' and finally 'no value' sentences. The macro average over the set of queries Q is then computed simply as:

$$NDGC(\mathbf{S}, k) = \frac{1}{|Q|} \sum_{j=1}^{|Q|} NDGC(S_j, k)$$

Note the similarity between the $rel(s_i)$ and the $val(s_i)$ used for the stratification purposes in Section 5.1. A question begs to be asked as to why a different scale is being used here. For the purposes of stratification I used the scale from 0 to 10. There it was important to introduce larger discrepancies between the more valuable and less valuable labels. By doing that I ensured that the richness was largely controlled by the more valuable labels instead of being dominated by the 'no value' and the 'potential value' sentences. The richness was used to compare the queries among themselves. I needed to prevent a situation where a query with many 'high value' and 'certain value' sentences would be deemed less rich than a query with considerably smaller number of such sentences. This could have easily occurred on the account of the first query having many 'no value' sentences and the second query having the 'potential value' label as the majority class. This problem does not exist for $rel(s_i)$ because the score is computed and normalized per query. Additionally, I am especially interested in detecting changes in the ranking of the sentences happening at the top of the list. Using the same scale as for $val(s_i)$ could interfere with the goal.

5.3 REPORTING AND STATISTICAL SIGNIFICANCE

For each method (or a set of methods) that I evaluate I generate the results following this procedure:

1. select one of the six (four for Chapter 6) folds as a test set,
2. if the method requires training use the remaining five (three for Chapter 6) folds,
3. if the training requires setting of hyper-parameters use part of the training set for validation,

4. apply the assessed ranking method to each query in the test set and order the sentences accordingly,
5. for each query use the sentence ordering to compute NDGC@10 and NDGC@100,
6. repeat the steps 1.–5. for the remaining folds.

The procedure yields two scores (NDGC@10 and NDGC@100) for each of the 42 queries. I report the unweighted means (i.e., the size of the query is not taken into account) of each score for the following groupings of the queries:

- Small sparse (SmSp) – 12 queries (e.g. cybercrime, nonindustrial use).
- Small dense (SmDs) – 18 queries (e.g. aural transfer, small manufacturer).
- Large sparse (LgSp) – 6 queries (e.g. identifying particular, essential step).
- Large dense (LgDs) – 6 queries (e.g. audiovisual work, fermented liquor).
- Overall (Overall) – This group consists of all the 42 queries.

Note that the groupings are based on the classification introduced for the purpose of stratified sampling described in Section 5.1. The unweighted means provide only a crude information about a method's performance. For this reason I also provide a box and whisker plot for the overall group and a scatter plot that distinguishes among the individual groups.

In [22] the author discusses statistical significance testing when k methods are applied to N data sets. Let c_i^j be the performance score of the j-th method on the i-th data set. The task is to decide whether, based on the values c_i^j, the methods are statistically significantly different and, in the case of more than two methods, which are the particular methods that differ in performance. The author assumes that variance is not recorded and nothing is assumed about the sampling scheme. The only requirement is that the measured results are "reliable" estimates of the method's performance on each data set. [22]

In my experiments, I treat each of the 42 queries as an individual data set and I use the NDGC@100 Overall as the evaluation metrics. Four sample plots of the scores distributions on different queries are shown in Figure 23. Although, it cannot be ruled out that these do follow the bell–shaped distribution required for the application of standard Gaussian statistics, it appears reasonable to assume they do not. Therefore, the use of a non-parametric test is appropriate. In [22] the Friedman test [36, 37], a non-parametric equivalent of the

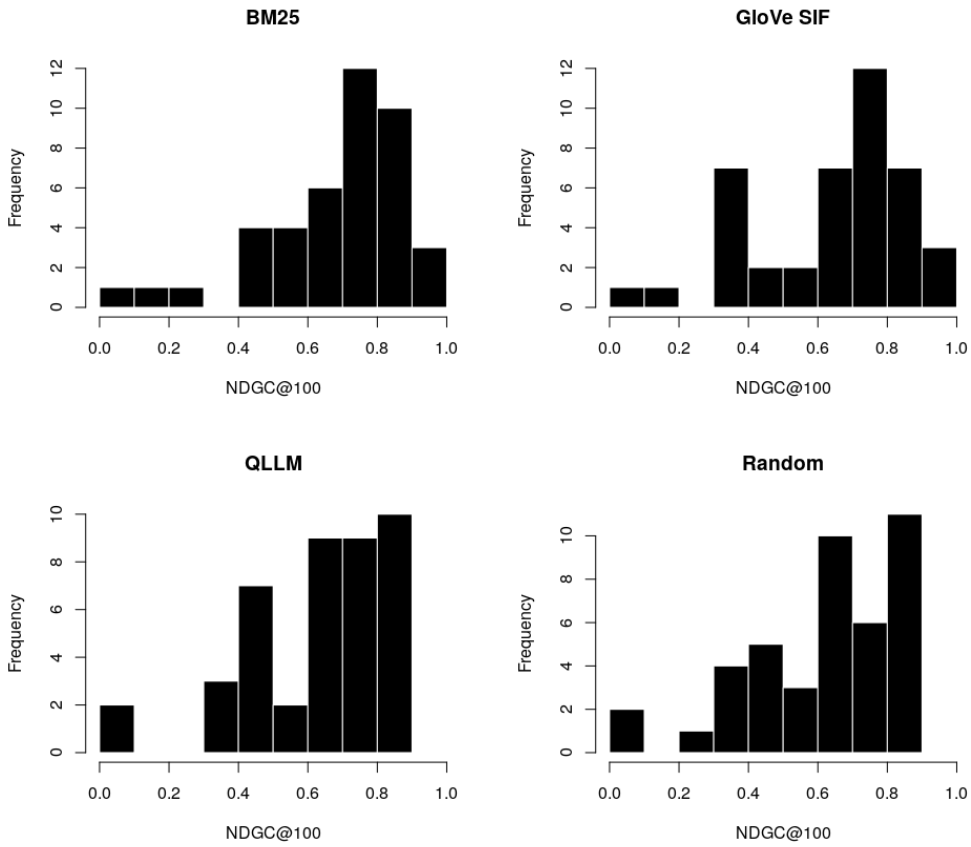

Figure 23: The figure shows four sample plots of the scores distributions on different queries.

repeated-measures ANOVA, is recommended. The test ranks the methods for each data set separately. In case of ties average ranks are assigned. The null-hypothesis states that all the methods are equivalent. Formally, let r_i^j be the rank of the j-th of k methods on the i-th of N data sets. The Friedman test compares the average ranks of methods computed as $R_j = \frac{1}{N}\sum_i r_i^j$ using the following statistic:

$$\chi_F^2 = \frac{12N}{k(k+1)}\left[\sum_j R_j^2 - \frac{k(k+1)^2}{4}\right]$$

The statistic is distributed according to χ_F^2 with $k-1$ degrees of freedom. [22]

The author of [22] points out that the authors of [54] showed that Friedman's χ_F^2 is

undesirably conservative and derived a better statistic:

$$F_F = \frac{(N-1)\chi_F^2}{N(k-1)-\chi_F^2}$$

The F_F statistic is distributed according to the F-distribution with $k-1$ and $(k-1)(N-1)$ degrees of freedom. [22] Therefore, I will use the F_F as recommended instead of the raw χ_F^2.

In case the null-hypothesis is rejected we can draw a conclusion that some methods do differ. In order to learn which of them are actually different a post-hoc test needs to be conducted. The subtlety of this step lies in the fact that it tests multiple hypotheses at the same time. In such a circumstance there is an increased danger of making a Type 1 error (false positive, i.e., rejection of a true null-hypothesis). Intuitively, if one tests many hypotheses at certain α there is a considerable chance that in one or more cases a null-hypothesis is going to be rejected purely by chance. This is a well-known statistical problem which has a general solution in the Bonferroni correction. [10] The author of [22] emphasizes that machine learning literature (e.g., [102]) have recognized that the Bonferroni correction is usually very conservative and weak since it supposes the independence of the hypotheses. Hence the use of other more powerful tests is recommended.

The Nemenyi test [82] is appropriate when all classifiers are compared to each other. A pair of methods differs significantly if their average ranks (R_j) differ by at least the critical difference:

$$CD = q_\alpha \sqrt{\frac{k(k+1)}{6N}}$$

The critical values q_α are based on the Studentized range statistic (infinite degrees of freedom) divided by $\sqrt{2}$. [22] Importantly, the Nemenyi test should not be used when we just want to compare multiple methods to a single control method. In such a situation even general procedures for controlling the family-wise error in multiple hypotheses testing are more appropriate. It is so because these adjust the critical value for making only $k-1$ comparisons as opposed to the Nemenyi test which anticiaptes $k(k-1)/2$ comparisons. The test statistic is then:

$$z = \frac{(R_i - R_j)}{\sqrt{\frac{k(k+1)}{6N}}}$$

In [22] the Bonnferoni-Dunn test [26] is suggested as a simple, yet more powerful, alternative for such a scenario. However, the author of [22] appears to advocate the use of either Holm's [50], Hochberg's [46], or Hommel's procedure [51] as even more powerful approaches.

Hochberg's and Hommel's methods reject more hypotheses than Holm's. Yet, they may exceed the prescribed family-wise error in certain circumstances. Since it has been reported that the differences between the enhanced methods are small [49] the author of [22] recommends the use of the Holm's method. [22] Therefore, in this work I employ the Holm's method as the post-hoc test. While the Nemenyi test establishes the critical value by dividing α with $k(k-1)/2$ the Bonferroni-Dunn test divides α with $k-1$. In case of a larger number of compared methods, say 10, the difference is considerable ($\alpha/45$ vs $\alpha/9$). The Holm's test is a step-down procedure which goes a little bit further. Here, the hypotheses are ordered by their significance from the most significant to the least. The p_i is then compared at $\alpha/(k-i)$. The comparisons are performed in the sequential order from the most significant hypotheses until a null-hypothesis that cannot be rejected is encountered. Then the procedure is stopped and it is assumed that none of the following null-hypotheses can be rejected. [22] It is quite easy to see how the Holm's test is more powerful than the Bonferroni-Dunn test for $i > 1$ provided the most significant null-hypothesis is rejected.

Finally, it is worth mentioning that a situation when the Friedmann test detects a statistically significant difference while a post-hoc test (i.e., the Holm's method in our case) fails to recognize even a single pair of methods that differ, may occur. This is due to the lower power of the post-hoc test. In such a situation it is still possible to conclude that some methods do differ but it is not possible to find out which. [22]

6.0 FEATURES INDICATIVE OF SENTENCE'S USEFULNESS

In a series of experiments I first explore performance of several conventional IR methods to document ranking when applied directly (Section 6.1). By direct application I mean a situation when a sentence is considered to be a document and the term of interest to be a query. I attempt to improve the ranking performance by informing the ranker with the sentence's context (Section 6.2). This is motivated by a well-known problem of the lack of overlap between a query and a short document such as a sentence. By taking the context into account I hope to mitigate this problem. Then, I assess effectiveness of a number of simple novelty detection methods, topic modelling techniques, and functional segmentation as to their filtering as well as ranking capabilities (Sections 6.3 and 6.4, and 6.5). The goal of these experiments is to get a detailed understanding of different aspects of the task. Furthermore, the results of the experiments provide strong indication as to how successfully could one tackle the task using traditional IR methods based mostly on measuring similarity between the query and the documents. Finally, I experiment with compound models where several of the above mentioned methods are integrated into a single more powerful approach (Section 6.6). The goal of this experiment is to ascertain that different methods do address different aspects of the task and, hence, the performance increases when they are applied in concert. All this together aims towards providing a solid basis for a compilation of a feature list that could be used in a classical learning-to-rank framework (Chapter 7). Finally, note that the experiments in this chapter are performed on four of the six folds described in Chapter 4. The second and the sixth folds are held out to be later used in Chapters 7 and 8.

6.1 RANKING SENTENCES DIRECTLY

In this Section I investigate the effectiveness of the direct approach to ranking, based on computing similarity between the query (the term of interest) and retrieved documents (sentences mentioning the term of interest), in the context of retrieving case law sentences for statutory interpretation. In general, I am interested if the similarity between a query (i.e., a statutory term) and a sentence somewhat captures the notion of sentence's usefulness for the interpretation. I look at the methods that consider only the terms that appear in both, the query and the sentence, as well as the methods that pay attention to all the terms in the sentence. I also assess the effectiveness of the hybrid methods combining both approaches.

6.1.1 Experiments

I first retrieve all the sentences that contain the term of interest. Using different strategies for measuring similarity of a sentence and a query, sentences are then ranked from the most similar to the least. The motivation is that the more similar the sentence is to the term of interest, the more likely it is about the term, and hence it may be useful for explaining its meaning. I experiment with different strategies of measuring similarity to analyze what impact they have on the performance. As a baseline I report the performance of a random system on a large sample of repeated runs. After each run I compute the average performance over all the preceding runs. I stop the iteration once the average remains stable for 10,000 consecutive runs. The performance of the random baseline is then the average performance to which the system converged.

I first evaluate an approach to measuring similarity of a query and a document based on the Okapi BM25 function, from the well know TF-IDF family. The function is defined as follows:

$$\text{BM25} = \sum_{t \in q} TF \cdot IDF \cdot QTF$$

$$\text{TF} = \frac{(k_1 + 1) \cdot tf_{td}}{k_1 \cdot \left(1 - b + b \cdot \frac{L_d}{L_{avg}}\right) + tf_{td}}$$

$$\text{IDF} = log\left(\frac{N - df_t + \frac{1}{2}}{df_t + \frac{1}{2}}\right) \quad \text{QTF} = \frac{(k_3 + 1) \cdot tf_{tq}}{k_3 + tf_{tq}}$$

Here, N is the size of the collection, df_t is the number of documents in which t occurs, tf_{td} is the frequency of term t in document d, tf_{tq} is the frequency of term t in query q, L_d and L_{avg} are the length of d and the average document length for the whole collection. k_1, k_3 and b are tuning parameters. [73, p. 233] For each of the four folds the parameters are optimized on the training set (the three remaining folds). For $k1$ and $k3$ I experimented with the range of 1.2–2.0. For b I tried the values between 0.0 and 1.0. In [79] the authors emphasize that BM25 is an example of a probabilistic approach to IR, [99] where the 2-Poisson model forms the basis for counting term frequency. [11, 45, 100] The idea is that there are two types of documents having term frequencies from different Poisson distributions—documents that are about a term and documents that only mention the term. The goal in IR is to distinguish between the two types. In order to achieve the goal the common version of BM25 considers query terms only, assuming that non-query terms are less useful for document ranking. [79]

The second approach I experiment with is a variant of TF-IDF tailored to accommodate some specifics of short documents such as sentences. The measure was presented in [2] and was subsequently referred to as TF-ISF [24, 30, 80]:

$$\text{TF-ISF} = \sum_{t \in q} log(tf_{td} + 1) \cdot log\left(\frac{N + 1}{\frac{1}{2} + df_t}\right) \cdot log(tf_{tq} + 1)$$

The three components of the formula are adapted versions of TF, IDF, and QTF. The meaning of the notation is the same as in the previous formula.

The language modeling approach to IR implements the idea that a document is a good match to a query if its language model is likely to generate the query. I work with a simple query likelihood language model presented in [89]:

$$P(d|q) \propto P(d) \prod_{t \in q} ((1-\lambda)P(t|M_c) + \lambda P(t|M_d))$$

Here, M_d is the language model of document d, M_c is a language model built from the entire document collection, and $0 < \lambda < 1$ is a hyperparameter. Correctly setting λ is important to the performance of this model. The equation captures the probability that the document that the user had in mind was d. [73, pp. 237–252] For each fold I optimized λ on the remaining 3 folds. Note that the language model, similarly to BM25 and TF-ISF, considers only the query terms as an evidence of a document relevance.

I also measure the similarity between projections of a query and a document in a low dimensional semantic space. In order to achieve this, I work with vector representations of words referred to as word embeddings. These representations are motivated by the so-called distributional hypothesis claiming that words that are used and occur in the same contexts tend to purport similar meanings [43]. The idea that "a word is characterized by the company it keeps" was popularized by Firth [33]. Regardless of a method by which word embeddings are generated, the idea is that the words with similar meanings are projected onto the vectors the cosine similarity of which is high. I use Gensim [98] to work with embeddings trained with various algorithms on a number of different corpora.

The use of word embeddings in IR has been explored quite extensively (see, e.g., [128, 79, 129, 39, 41]). The main appeal of word embeddings in the context of IR lies in their ability to measure similarity between an arbitrary pair of words (w_i, w_j) by, e.g., measuring the cosine similarity between their word embedding vectors ($\vec{w}_i$, $\vec{w}_j$):

$$sim(w_i, w_j) = cos(\vec{w}_i, \vec{w}_j) = \frac{\vec{w}_i^T \vec{w}_j}{\|\vec{w}_i\| \|\vec{w}_j\|}$$

Therefore, instead of considering only the query terms repetition as evidence of aboutness it is possible to consider the relationship between the query terms and all the terms in a document. [79] This makes the approach an appealing solution to a well-known mismatch problem. [128, 129, 39] The problem refers to the situation when a relevant document does not have any (or enough) terms in common with the query. In this work I only consider sentences that are guaranteed to contain at least one match of the query. Hence, the mismatch problem

is not a concern. However, the opportunity to measure similarity between an arbitrary pair of words is still appealing. Intuitively, the methods that are constrained to operate on query terms only have quite a limited capability of distinguishing between relevant and not relevant documents. As mentioned before, this problem is especially pronounced in case of short documents such as sentences. It appears that something different, from the query terms overlap, needs to be the basis for the relevance assessment in this work. The close relatedness of the other terms in the sentence to the terms contained in the query seems to be a viable candidate.

From the numerous available options I chose to experiment with word embeddings generated by the word2vec algorithm based on the skip-gram model [77, 76], the GloVe model that combines global matrix factorization and local context window methods [87], and the FastText algorithm based on the skip-gram model, where each word is represented as a bag of character n-grams [9, 57, 56]. Furthermore, I use all the 538,680,570 sentences from the case law database (Chapter 4) to train domain specific embeddings. To train the embeddings I used the sequences of lemmatized tokens in lower case as described in Chapter 4. Before training the embeddings I trained a model for detection of 2- and 3-word long phrases to be included in the embeddings' training.[1] [77, 12] I trained embeddings with 300 dimensions and finished the training after 10 iterations over the corpus.[2]

Word embeddings are a convenient way to express the meaning of individual words or common multi-word expressions (e.g., New York, computer program). Interestingly, how should one go about combining the word vectors, for the purpose of representing the semantics of word sequences (phrases, sentences, paragraphs, or whole documents), is not straightforward. The simplest approach is to compute a so-called document centroid which is just an average embedding of the words contained in the document:

$$\vec{d} = \frac{1}{|d|} \sum_{w_j \in d} \frac{\vec{w}_j}{\|\vec{w}_j\|}$$

In this work I use this approach because of its simplicity. It has been shown that the

[1] I used Gensim's phrases module which is available at https://radimrehurek.com/gensim/models/phrases.html

[2] See https://radimrehurek.com/gensim/models/word2vec.html.

method does not work very well in many contexts. [123] More sophisticated methods such as convolutional neural networks and recurrent neural networks to compute word sequence embeddings have been proposed in recent work. [55, 64, 60, 111, 59, 113, 122] However, it appears that much simpler weakly supervised [123] or completely unsupervised [3] methods achieve competitive performance on a variety of tasks. Therefore, in addition to simple averages I also work with document vectors where the words are weighted with the smooth inverse frequency (SIF) as proposed in [3]. Using SIF a sentence embedding is computed as a weighted average of the word vectors. Then a projection of a set of sentences on their first singular vector is removed (common component removal). [3] The weight of a word w is:

$$w = \frac{a}{a + p(w)}$$

Here a is a parameter and $p(w)$ the (estimated) word frequency. [3]

To compute the similarity both the query $\mathbf{q} = \{t_1, t_2, \dots\}$ and the document $\mathbf{d} = \{t_1, t_2, \dots\}$ are first projected onto the vectors $\vec{q}$ and $\vec{d}$. The distance between $\vec{q}$ and $\vec{d}$ can be determined by measuring the cosine of the angle between the two vectors:

$$sim(\mathbf{q}, \mathbf{d}) = cos(\vec{q}, \vec{d}) = \frac{\vec{q}^T \vec{d}}{\|\vec{q}\| \|\vec{d}\|}$$

Here, $\vec{q}$ is a compound vector over $\vec{t_i}$ in query $\mathbf{q}$ and $\vec{d}$ is a compound vector over $\vec{t_i}$ in document $\mathbf{d}$. I use the word2vec model (300 dimensions) trained on the Google News corpus (about 100 billion words)[3] [77, 76, 78], the GloVe model (300 dimensions) trained on the Wikipedia 2014 and Gigaword 5 corpora (6 billion tokens)[4] [87],[5] and FastText model (300 dimensions) trained on Wikipedia,[6] [9].

I used the embeddings trained on the court decisions corpus (Chapter 4) to experiment with the so-called dual embedding space model (DESM) described in [79]. The working of this model is quite similar to the simple computation of similarity between the centroid vectors of a query and a document. There are two important differences:

[3] https://code.google.com/archive/p/word2vec/
[4] https://nlp.stanford.edu/projects/glove/
[5] Both available at https://github.com/RaRe-Technologies/gensim-data
[6] https://github.com/facebookresearch/fastText/blob/master/pretrained-vectors.md

court		offense		independent	
IN-IN	IN-OUT	IN-IN	IN-OUT	IN-IN	IN-OUT
court	supreme	offense	commit	independent	contractorship
judge	family	crime	extraneous	independently	contractor
chancellor	surrogate	felony	sex	distinct	independent
tribunal	county	offence	lesser	separate	entirely
referee	justice	burglary	nonsex	derivative	wholly
we	marine	murder	jailable	autonomous	restoration

Table 9: The table shows the different results of retrieving the words that are the most similar to "court," "offense," and "independent" in the IN-IN and IN-OUT space based on the word embeddings model trained on the court decisions corpus.

- Two jointly learned embedding models are used—one for the query and one for the document.
- Each single word from the query is compared to the document centroid separately and the measured similarities over the query words are aggregated.

The authors of [79] emphasize: A crucial detail often overlooked when using Word2Vec is that there are two different sets of vectors referred to as IN and OUT. By default, Word2Vec discards OUT vectors at the end of training and outputs only IN vectors. The interactions of the IN and OUT word embeddings provide additional distributional semantics that are not observable by considering any of the two embeddings in isolation. [79] While the typically used IN-IN cosine similarities are higher for words that are synonymous or similar, the IN-OUT cosine similarities are higher for words that often co-occur. Table 9 shows the different results of retrieving the words that are the most similar to "court," "offense," and "independent" in the IN-IN and IN-OUT space.

The authors of [79] present two variants of DESM corresponding to retrieval in the IN-OUT space and the IN-IN space:

$$DESM_{IN-OUT}(\mathbf{q}, \mathbf{d}) = \frac{1}{|\mathbf{q}|} \sum_{\vec{q}_i \in \mathbf{q}} \frac{\vec{q}_{IN,i}^T \vec{d}_{OUT}}{\|\vec{q}_{IN,i}\| \|\vec{d}_{OUT}\|}$$

$$DESM_{IN-IN}(\mathbf{q}, \mathbf{d}) = \frac{1}{|\mathbf{q}|} \sum_{\vec{q}_i \in \mathbf{q}} \frac{\vec{q}_{IN,i}^T \vec{d}_{IN}}{\|\vec{q}_{IN,i}\| \|\vec{d}_{IN}\|}$$

In [79] the authors show that $DESM_{IN-OUT}$ (topical similarity) is a better indication of aboutness than $DESM_{IN-IN}$ (typical similarity). However, they also present evidence that $DESM$ is a weak general ranker and that it is effective only at ranking at high positions (telescoped settings). Therefore, they propose a mixture model combining $DESM$ with a term frequency based measure, such as BM25. The mixture model is defined as:

$$MM(\mathbf{q}, \mathbf{d}) = \alpha DESM(\mathbf{q}, \mathbf{d}) + (1 - \alpha) BM25(\mathbf{q}, \mathbf{d})$$

The authors optimized α considering values between 0 and 1 at 0.01 steps. They do not report which α turned out to work best in their experiments. [79]

6.1.2 Results and Discussion

The results of the experiments described in Section 6.1.1 are reported in Table 10 (group and overall means) and Figure 24 (box and whisker plots and swarm plots). On the new larger data set (28 queries) I confirm the results from [108] (three queries). Overall, the new results indicate that the direct approach does not appear to be very effective. In [108] of all the six tested methods only TF-ISF outperformed the Random baseline by larger than a negligible margin. Here, BM25 is another method that appears to improve over the Random baseline. Note that both methods are based on considering the repetition of the query terms only. The mixed methods (MM) based on the BM25 do not appear to improve over it.

First, I examine the thesis statement S1.1 by testing the corresponding null hypothesis. As the first step I use the Friedman test to assess the hypothesis that at least one of the Random, BM25, TF-ISF, and QLLM methods differs from the other three. The methods are compared in terms of their performance as measured with NDGC@100 on the Overall. I reject the null hypothesis (p-value=.0004) and conclude that there is at least one pair

Method	SmSp @10	SmSp @100	SmDs @10	SmDs @100	LgSp @10	LgSp @100	LgDs @10	LgDs @100	Overall @10	Overall @100
Random	.40 ± .07	.69 ± .15	.52 ± .08	.76 ± .11	.29 ± .16	.35 ± .18	.47 ± .11	.47 ± .11	.45 ± .12	.64 ± .20
BM25	.50 ± .14	.75 ± .12	.62 ± .17	.80 ± .12	.46 ± .07	.49 ± .15	.64 ± .19	.58 ± .12	.57 ± .16	**.71 ± .17**
TF-ISF	.47 ± .13	.73 ± .13	.62 ± .16	.80 ± .12	.45 ± .17	.51 ± .13	.61 ± .17	.57 ± .07	.55 ± .17	**.71 ± .16**
QLLM	.46 ± .09	.71 ± .13	.54 ± .05	.76 ± .09	.27 ± .18	.34 ± .19	.57 ± .20	.48 ± .13	.48 ± .14	.65 ± .20
GloVe	.47 ± .17	.73 ± .15	.49 ± .14	.77 ± .09	.27 ± .04	.37 ± .07	.33 ± .28	.47 ± .20	.43 ± .18	.66 ± .20
GloVe^{SIF}	.53 ± .23	.74 ± .18	.49 ± .13	.78 ± .08	.19 ± .10	.32 ± .13	.33 ± .27	.48 ± .17	.44 ± .21	.66 ± .22
w2vec	.46 ± .16	.73 ± .14	.50 ± .09	.77 ± .09	.16 ± .06	.39 ± .05	.35 ± .30	.49 ± .19	.42 ± .18	.66 ± .19
w2vec^{SIF}	.47 ± .21	.72 ± .14	.47 ± .14	.76 ± .09	.15 ± .10	.36 ± .07	.31 ± .30	.49 ± .18	.40 ± .21	.65 ± .19
fastt	.45 ± .18	.74 ± .13	.50 ± .12	.77 ± .09	.19 ± .03	.35 ± .08	.33 ± .29	.49 ± .22	.42 ± .19	.66 ± .20
fastt^{SIF}	.50 ± .19	.73 ± .15	.51 ± .16	.78 ± .08	.17 ± .10	.37 ± .09	.32 ± .31	.48 ± .20	.43 ± .22	.66 ± .20
DESM$_{I\times I}$	.46 ± .13	.72 ± .14	.51 ± .11	.77 ± .08	.26 ± .08	.38 ± .10	.39 ± .35	.53 ± .22	.44 ± .18	.67 ± .19
DESM$_{I\times O}$	.46 ± .17	.72 ± .16	.48 ± .09	.76 ± .08	.11 ± .06	.32 ± .12	.42 ± .31	.50 ± .20	.41 ± .19	.65 ± .20
MM$_{I\times I}$	.50 ± .13	.75 ± .12	.61 ± .15	.80 ± .11	.49 ± .08	.50 ± .13	.62 ± .23	.57 ± .15	.57 ± .15	.71 ± .17
MM$_{I\times O}$	.50 ± .14	.75 ± .12	.62 ± .17	.80 ± .12	.47 ± .08	.49 ± .15	.64 ± .19	.58 ± .12	.57 ± .16	.71 ± .17

Table 10: The table shows the results of the experiments on direct retrieval of sentences. The NDGC@10 and NDGC@100 are shown for the small sparse queries (SmSp), small dense queries (SmDs), large sparse queries (LgSp), large dense queries (LgDs), and all of them together (Overall).

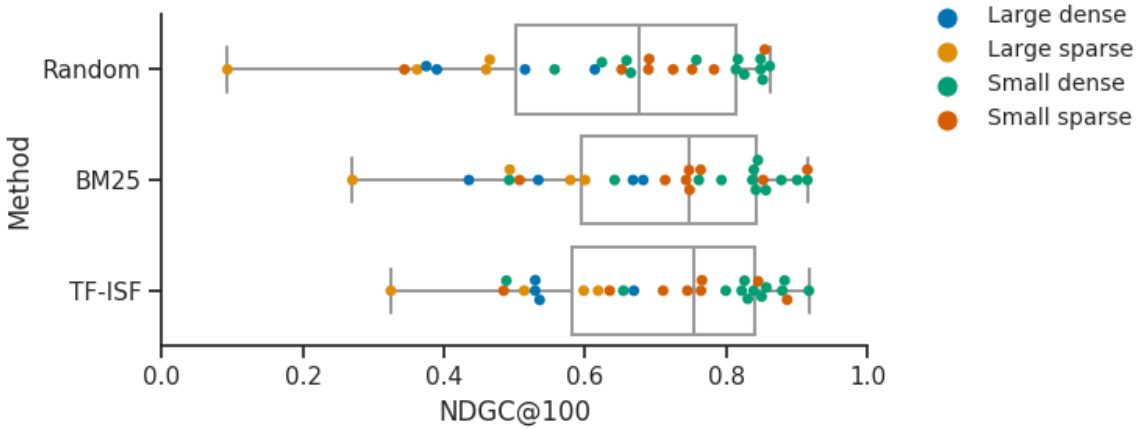

Figure 24: The figure shows scatter plots of the performance on the individual 28 queries measured in terms of NDGC@100. The two methods (BM25 and TF-ISF) that outperform the Random baseline are shown.

of methods that differ from each other. Since I am interested if any of the methods is different from the Random baseline I use the Holm procedure as described in Subsection

5.3. BM25 has the lowest p-value (.003) when tested against the Random. When I adjust the p-value for 3 comparisons I reject the null for BM25 (p-value=.008). Since I rejected the null hypothesis for BM25 I can step down to test TF-ISF, the method with the second lowest p-value (.004). After applying the adjustment ($\alpha/(k-2)$) I reject the null for TF-ISF (p-value=.007). Finally, I test the QLLM method (p-value=.3) for which I cannot reject the null. Hence, I have established that the BM25 and TF-ISF methods are significantly different from Random.

Second, I analyze the thesis statement S1.2 by testing the corresponding null hypothesis. Again, as the first step I use the Friedman test to assess the hypothesis that at least one of the Random, GloVe, GloVe^{SIF}, w2vec, w2vec^{SIF}, fastt, fastt^{SIF}, DESM$^{I\times I}$, and DESM$^{I\times O}$ methods differs from the other ones. Interestingly, I fail to reject the null hypothesis (p-value=.763). Hence, I conclude that rankings produced by the selected methods do not differ from ranking the sentences randomly.

Finally, I would like to look at the thesis statement S1.3. I selected the BM25 method as a baseline since it appears to be the strongest single paradigm method I have assessed so far. Hence, I use the Friedmann test to test the hypothesis that at least one of the BM25, $MM_{I\times I}$, and $MM_{I\times O}$ methods differs from the other ones. Again, I fail to reject the null hypothesis (p-value=.921).

Overall, it appears that the direct approach to sentence retrieval does not work very well. This conclusion is in agreement with the conclusion we reached in [108]. There TF-ISF was the only method that seemed to outperform the Random baseline. Here, BM25 shows similar promise. Leaving aside the mixed models ($MM_{I\times I}$ and $MM_{I\times O}$), all the other methods perform similarly to Random. In [108] we claimed that the root cause of the low performance, in the context of this specific task, appears to be the preference of the systems for short over long sentences. For example, the following sentences top the rankings produced by most of the models:

i. The “Aural transfer.”
[term of interest: “aural transfer;” gold label: ‘no value’]

ii. Here, plaintiff made an “aural transfer.”
[term of interest: “aural transfer;” gold label: ‘potential value’]

iii. An “aural transfer” means a transfer containing the human voice at any point.
[term of interest: “aural transfer;” gold label: ‘high value’]

From these sentences only iii. has 'high value' whereas the rest have 'no value' or 'potential value.' However, this makes sense because, strictly speaking, the most similar sentence to a query is going to be the one which contains exactly the same terms as the query. Since the provision of additional valuable information is part of the relevance definition in this work the preference for short sentences is not a good strategy. That is one of the reasons why the methods that are primarily meant to measure similarity between word sequences barely outperform the Random baseline.

The relative success of the TF-ISF method could be explained by the deliberate omission of the normalization based on a document length in the TF part. Unlike other studied methods, TF-ISF prefers longer sentences that mention the terms from the query multiple times over the shorter sentences presented above. It appears this strategy is clearly preferable in sentence retrieval for statutory interpretation. Unlike in [108], where we used recommended values of $k_1, k_3 = 1.2$ and $b = 0.75$ for the BM25 method, here I optimized the parameters on the training set. For all of the folds these settled on $k_1 = 2.0$ (the highest value I allowed) and $b = 0.1$ (from 0 to 1). Note that k_3 does not play any role under these settings. Setting k_1 high and b low makes the measure quite similar in behavior to the TF-ISF method. With parameters thus set BM25 also prefers sentences that contain multiple occurrences of the query terms.

Unlike in [79] I did not observe mixed models (linear combination of BM25 and $DESM_{I\times I}$ or $DESM_{I\times O}$) outperforming the base BM25 method. I believe that this is due to a more complex notion of relevance that I employ in sentence retrieval for statutory interpretation when compared to a typical notion employed in general web search (such as in [79]). For example, when the source provision is a definition of the term of interest then a retrieved sentence that is a verbatim copy of the source provision is about the term as much as it can be. In the standard IR settings such a sentence should be ranked very high. Yet, in the context of this work the sentence has 'no value.' There is no information that is being passed to the methods that could be used to tackle this problem. Another challenging piece of the relevance definition is the difference in the meaning of the terms. What I consider as term used in a different meaning (because it is not useful for the interpretation of the term of interest) would often pass as the same meaning in general IR (e.g., 'useful improvement'

in an invention to qualify for a patent vs 'useful improvement' in general sense).

Finally, it is worth mentioning that retrieval of short texts such as sentences has been considered as quite challenging even in general IR. It appears to be well-established that the traditional approach to ranking, based on computing similarity between the query and retrieved documents, is less effective for very short documents such as sentences (see, e.g., [81]). Usually the main cause of the decreased performance is the lack of a robust overlap between a query and a sentence. The similarity between the query and a document is based either on a direct overlap of terms or on an "overlap" in their meanings. In larger documents an unusually high occurrence of a term or its meaning strongly indicates that the document might be "about" the term. This "aboutness" assumption often works surprisingly well for longer documents. In shorter documents there is typically just one exact occurrence of the term. Even in case of more than one occurrence it is questionable if the "aboutness" assumption would still be as solid as for longer documents. As a result, there are no grounds on which the short documents could be ranked reliably based solely on the term occurrence statistics. This is why methods capable of matching the query terms to the whole sentence appeared so promising. For many reasons, some of which were discussed above, these do not appear to be helpful for the specific task of retrieving sentences for statutory interpretation.

Even though I was able to show that the BM25 and the TF-ISF methods do outperform the Random baseline I conclude that the direct approach to sentence retrieval does not work well. First, the Random is a very weak baseline—typically the weakest one there is. Second, if one considers how do the two successful methods work (making the relevance assessment based almost exclusively on the frequency of the query terms) in light of the definition of the task of retrieving sentences for statutory interpretation (Chapter 3) it appears very likely that there is ample space for improvement.

6.2 SMOOTHING SENTENCES WITH CONTEXT

In this section I analyze the effects of considering the context of a sentence when deciding about its ranking. The motivation behind this approach is that parts of a decision other

than the sentence itself could be suggestive about the value of the sentence. For example, if I determine that the whole decision is "about" the term of interest then I expect that at least some of the sentences it contains will be "about" the term—and these are the sentences that I would like to rank highly.

6.2.1 Experiments

As in case of the previous batch of experiments (Section 6.1), I first retrieve all the sentences that contain the term of interest. Unlike in the first batch (Section 6.1) I also retrieve the remaining parts of the decisions from which the sentences come. Using some of the similarity measuring strategies as before I ranked the sentences based on their similarity to the query as well as the similarity of their varying contexts to the query.

The first approach I experiment with to incorporate context into the similarity measurement is linear interpolation applied to similarities as measured on the sentence itself and on a fixed context. The types of context are the whole case, an opinion, and a paragraph (as described in Section 4). I introduce the hyperparameter λ_1 that controls the weighting of the two pieces (the sentence itself and the context). The general form of this approach is then:

$$Sim\text{-}i = (1 - \lambda_1)Sim_s + \lambda_1 Sim_i$$

Here, Sim_s is the similarity as measured between the sentence and the query, Sim_i is the similarity as measured between the context and the query where:

$$i \in \{\text{c(ase)}, \text{o(pinion)}, \text{p(aragraph)}\}$$

λ_1 controls the weight assigned to each component (the higher λ_1 the more importance the context is given). Setting λ_1 to 0.0 makes the method equal to its counterpart that does not take context into account (Subsection 6.1). I optimize λ_1 on the training set.

The general form has slightly varying implementations across the tested similarity functions. For both BM25 the implementation is as follows:

$$\text{BM25-i} = \sum_{t \in q} [(1 - \lambda_1) TF_s \cdot IDF_s + \lambda_1 TF_i \cdot IDF_i] \cdot QTF$$

Note that I do not test TF-ISF since BM25 can be optimized to work very similarly (see Section 6.1). The QLLM already had one hyperparameter (λ) in its original form. Thus, the context-aware form needs to accommodate two hyperparameters where λ_2 stands for the original λ:

$$P(d|q) \propto P(d) \prod_{t \in q} [(1 - \lambda_1 - \lambda_2) P(t|M_c) + \lambda_1 P(t|M_i) + \lambda_2 P(t|M_d)]$$

The new M_i element is the language model of context i. The interpolated cosine similarity for representations based on word embeddings is straightforward:

$$\text{COS-i} = (1 - \lambda_1) COS_s + \lambda_1 COS_i$$

For the $DESM_{IN-IN}$ and $DESM_{IN-OUT}$ methods the linear interpolation has the following form:

$$DESM_{IN-IN}(\mathbf{q}, \mathbf{d}, \mathbf{c}) = (1 - \lambda_1) \frac{1}{|\mathbf{q}|} \sum_{\vec{q}_i \in \mathbf{q}} \frac{\vec{q}_{IN,i}^T \vec{d}_{IN}}{\|\vec{q}_{IN,i}\| \|\vec{c}_{IN}\|} + \lambda \frac{1}{|\mathbf{q}|} \sum_{\vec{q}_i \in \mathbf{q}} \frac{\vec{q}_{IN,i}^T \vec{c}_{IN}}{\|\vec{q}_{IN,i}\| \|\vec{c}_{IN}\|}$$

The second approach I evaluate is the recursive interpolation presented in [24] originally designed for TF-ISF. The idea is to consider n preceding and n following sentences as the context of the sentence I wish to evaluate. The sentences that are closer to the focused sentence are assigned more weight than those that are further away. The general formula for this approach is the following:

$$Sim\text{-}r_s = (1 - \lambda_1) Sim_s + \lambda_1 [Sim\text{-}r_{s-} + Sim\text{-}r_{s+}]$$

Here, Sim_s is the similarity score computed between the focused sentence and the query, $s-$ and $s+$ stand for the preceding and the next sentence respectively, and λ_1 is the hyperparameter controlling the weight assigned to the context. The authors of [24] set n to 3 (i.e., the size of the context is the 3 preceding and the 3 next sentences) suggesting that it could

be extended to the whole document if necessary. However, I found that for $n > 4$ the computation becomes expensive. Since I observed increased performance with larger n on the development set I set n to 4. The implementation for the specific methods is straightforward where Sim_s is replaced with the respective method.

Finally, as I followed the suggestion presented in [79] in Section 6.1 here I also experiment with mixture models. These combine the methods that match the query terms only with methods that match all the words in a document. Based on the experiments I selected BM25-p, i.e., the BM25 method smoothed with the paragraph context, as the most promising 'only query terms' method. Similarly, I selected fastt-c^{SIF} (fastt^{SIF} with the case context) and GloVe-r (GloVe with recursive context) as the most promissing methods that consider the whole document.

6.2.2 Results and Discussion

The results of the experiments described in Section 6.2.1 are reported in Table 11 (group and overall means) and Figure 25 (box and whisker plots and swarm plots). In [108] application of the context-aware methods led to a major improvement. Here, it also appears that some improvement is achieved but it does not seem that dramatic. Nevertheless, the mean NDGC scores reported in Table 11 suggest that it is advantageous to take context into account. This holds for many variants of the context-aware methods from the both categories. As in case of the direct approach to ranking (Section 6.1) it does not appear that the mixed models work much better than the individual ones.

First, I examine the thesis statement S2.1 by testing the corresponding null hypothesis. As the first step I use the Friedman test to assess the hypothesis that at least one of the BM25, BM25-c, BM25-o, BM25-p, BM25-r, QLLM-c, QLLM-o, QLLM-p, and QLLM-r methods differs from the other eight. The methods are compared in terms of their performance as measured with NDGC@100 on the Overall. I reject the null hypothesis (p-value=.00001) and conclude that there is at least one pair of methods that differ from each other. Since I am interested if any of the methods is different from the plain BM25 baseline I use the Holm procedure as described in Subsection 5.3. QLLM-p has the lowest p-value (.001) when

Method	SmSp @10	SmSp @100	SmDs @10	SmDs @100	LgSp @10	LgSp @100	LgDs @10	LgDs @100	Overall @10	Overall @100
BM25	.50 ± .14	.75 ± .12	.62 ± .17	.80 ± .12	.46 ± .07	.49 ± .15	.64 ± .19	.58 ± .12	.57 ± .16	.71 ± .17
$\text{DESM}_{I\times I}$	.46 ± .13	.72 ± .14	.51 ± .11	.77 ± .08	.26 ± .08	.38 ± .10	.39 ± .35	.53 ± .22	.44 ± .18	.67 ± .19
BM25-c	.47 ± .14	.76 ± .10	.62 ± .17	.80 ± .12	.51 ± .06	.51 ± .12	.67 ± .22	.56 ± .14	.57 ± .17	.71 ± .16
BM25-o	.46 ± .17	.76 ± .11	.61 ± .17	.80 ± .12	.51 ± .06	.51 ± .10	.63 ± .22	.55 ± .16	.56 ± .17	.71 ± .16
BM25-p	.51 ± .12	.74 ± .13	.64 ± .14	.81 ± .11	.59 ± .16	.54 ± .05	.64 ± .27	.59 ± .18	.60 ± .16	.72 ± .16
BM25-r	.43 ± .17	.75 ± .12	.54 ± .14	.78 ± .11	.51 ± .01	.50 ± .08	.64 ± .13	.56 ± .10	.52 ± .15	.70 ± .15
QLLM-c	.44 ± .10	.72 ± .12	.53 ± .09	.79 ± .10	.51 ± .09	.54 ± .07	.55 ± .21	.51 ± .21	.50 ± .12	.69 ± .16
QLLM-o	.44 ± .14	.73 ± .12	.52 ± .12	.78 ± .11	.43 ± .04	.52 ± .05	.59 ± .20	.50 ± .18	.49 ± .14	.69 ± .16
QLLM-p	.46 ± .09	.71 ± .13	.54 ± .05	.76 ± .09	.27 ± .18	.34 ± .19	.57 ± .20	.48 ± .13	.48 ± .14	.65 ± .20
QLLM-r	.42 ± .14	.71 ± .11	.54 ± .06	.78 ± .10	.49 ± .11	.55 ± .06	.48 ± .29	.48 ± .18	.49 ± .14	.68 ± .15
GloVe-c	.48 ± .16	.74 ± .11	.52 ± .08	.78 ± .10	.35 ± .28	.43 ± .12	.39 ± .18	.49 ± .20	.46 ± .16	.68 ± .18
GloVe-o	.47 ± .19	.74 ± .13	.57 ± .09	.79 ± .09	.37 ± .18	.42 ± .11	.35 ± .22	.46 ± .23	.48 ± .17	.68 ± .20
GloVe-p	.44 ± .15	.73 ± .14	.50 ± .13	.76 ± .10	.36 ± .10	.38 ± .10	.36 ± .28	.51 ± .19	.44 ± .16	.66 ± .19
GloVe-r	.42 ± .15	.75 ± .09	.56 ± .07	.79 ± .10	.58 ± .23	.59 ± .09	.45 ± .35	.50 ± .25	.51 ± .18	.71 ± .16
fastt-c^{SIF}	.44 ± .11	.72 ± .13	.56 ± .10	.79 ± .11	.49 ± .09	.48 ± .08	.48 ± .13	.57 ± .15	.50 ± .11	.69 ± .16
fastt-o^{SIF}	.44 ± .12	.71 ± .14	.55 ± .12	.79 ± .12	.40 ± .18	.47 ± .10	.59 ± .20	.54 ± .19	.50 ± .15	.69 ± .18
fastt-p^{SIF}	.37 ± .11	.68 ± .17	.53 ± .16	.77 ± .13	.31 ± .12	.37 ± .14	.55 ± .32	.49 ± .25	.45 ± .19	.65 ± .21
fastt-r^{SIF}	.51 ± .17	.76 ± .11	.55 ± .15	.77 ± .09	.47 ± .20	.52 ± .06	.35 ± .40	.48 ± .25	.50 ± .21	.69 ± .17
$\text{DESM-c}_{I\times I}$	.46 ± .11	.73 ± .13	.56 ± .10	.80 ± .08	.19 ± .22	.38 ± .18	.42 ± .30	.51 ± .22	.46 ± .19	.68 ± .20
$\text{DESM-o}_{I\times I}$	.49 ± .14	.74 ± .15	.54 ± .12	.78 ± .10	.29 ± .26	.39 ± .22	.36 ± .26	.51 ± .20	.46 ± .18	.68 ± .21
$\text{DESM-p}_{I\times I}$	.43 ± .10	.71 ± .13	.52 ± .10	.77 ± .10	.34 ± .11	.37 ± .11	.35 ± .34	.51 ± .23	.45 ± .16	.66 ± .19
$\text{DESM-r}_{I\times I}$	.51 ± .17	.76 ± .11	.55 ± .15	.77 ± .09	.47 ± .20	.52 ± .06	.35 ± .40	.48 ± .25	.50 ± .21	.69 ± .17
$\text{DESM-c}_{I\times O}$	.46 ± .16	.72 ± .15	.56 ± .12	.79 ± .09	.22 ± .22	.37 ± .17	.58 ± .23	.57 ± .16	.48 ± .19	.68 ± .19
$\text{DESM-o}_{I\times O}$	.45 ± .15	.72 ± .14	.54 ± .12	.79 ± .09	.21 ± .18	.36 ± .15	.53 ± .30	.54 ± .18	.47 ± .20	.67 ± .19
$\text{DESM-p}_{I\times O}$	.41 ± .13	.70 ± .16	.45 ± .11	.75 ± .11	.15 ± .07	.34 ± .13	.44 ± .24	.48 ± .21	.40 ± .16	.64 ± .20
$\text{DESM-r}_{I\times O}$	.47 ± .16	.74 ± .10	.53 ± .15	.76 ± .09	.44 ± .20	.46 ± .19	.38 ± .29	.46 ± .25	.48 ± .18	.67 ± .19
BM-p+fastt-c	.49 ± .13	.73 ± .12	.62 ± .15	.81 ± .10	.60 ± .17	.57 ± .05	.59 ± .25	.59 ± .19	.58 ± .16	.72 ± .15
BM-p+GlV-r	.52 ± .13	.75 ± .12	.63 ± .16	.80 ± .11	.63 ± .18	.57 ± .05	.54 ± .34	.57 ± .21	.58 ± .18	.72 ± .15

Table 11: The table shows the results of the experiments on smoothing sentences with their context. The NDGC@10 and NDGC@100 are shown for the small sparse queries (SmSp), small dense queries (SmDs), large sparse queries (LgSp), large dense queries (LgDs), and all of them together (Overall).

tested against the BM25. When I adjust the p-value for 8 comparisons I reject the null for QLLM-p (p-value=.004). Since I rejected the null hypothesis for QLLM-p I can step down to test QLLM-r, the method with the second lowest p-value (.005). For this method I reject the null as well (p-value=.03). For QLLM-o, the method with the next lowest p-value (.008), I cannot reject the null (p-value=.0504). Hence, I have been only able to establish that the QLLM-p and QLLM-r methods are significantly different (weaker performance) from BM25.

Second, I analyze the thesis statement S2.2 by testing the corresponding null hypothesis. Again, as the first step I use the Friedman test to assess the hypothesis that at least

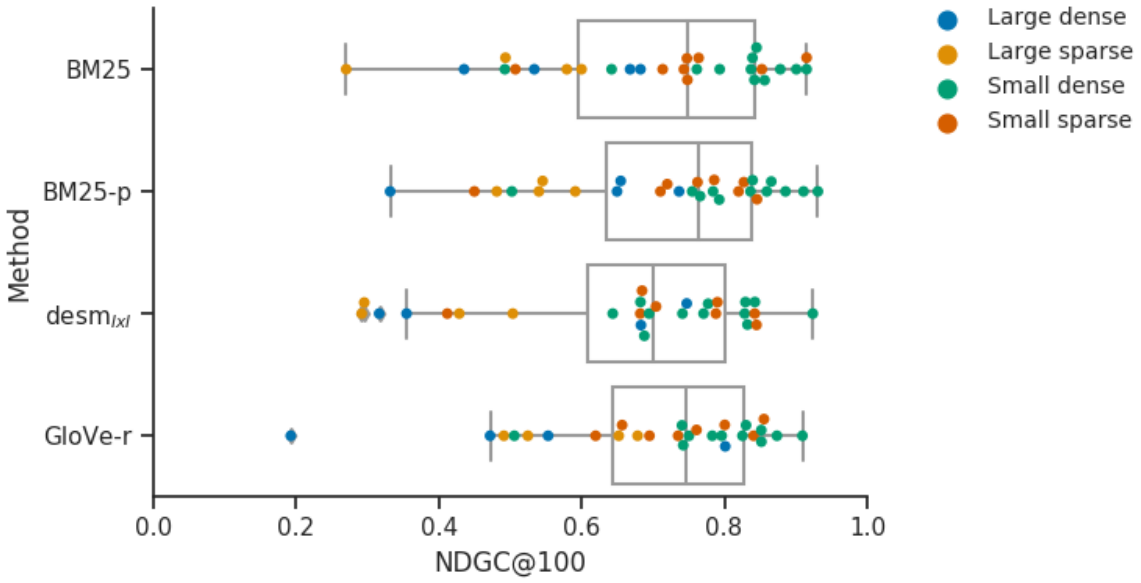

Figure 25: The figure shows scatter plots of the performance on the individual 28 queries measured in terms of NDGC@100. Both BM25-p and GloVe-r appear to perform slightly better than the best similar methods that do not use context (BM25 and DESM$_{I\times I}$).

one of the fastt^{SIF}, GloVe-c, GloVe-o, GloVe-p, GloVe-r, fastt-c^{SIF}, fastt-o^{SIF}, fastt-p^{SIF}, fastt-r^{SIF}, DESM-c$_{I\times I}$, DESM-o$_{I\times I}$, DESM-p$_{I\times I}$, DESM-r$_{I\times I}$, DESM-c$_{I\times O}$, DESM-o$_{I\times O}$, DESM-p$_{I\times O}$, and DESM-r$_{I\times O}$ methods differs from the other ones. I reject the null hypothesis (p-value=.047) and conclude that there is at least one pair of methods that differ from each other. The fastt-c^{SIF} method has the lowest p-value (.01) when tested against fastt^{SIF}. When I adjust the p-value for 16 comparisons I cannot reject the null for fastt-c^{SIF} (p-value=.183). Hence, I conclude that rankings produced by the selected context-aware methods do not differ from rankings produced by the methods operating on the sentences only.

Finally, I would like to look at the thesis statement S2.3. I selected the BM25-p method as a baseline since it appears to be the strongest single paradigm method I have assessed so far. Hence, I use the Friedman test to assess the hypothesis that at least one of the BM25-p, BM-p+fastt-c, and BM-p+GlV-r methods differs from the other ones. Again, we fail to

	BM25	QLLM	GloVe	fastt^{SIF}	DESM$_{I\times I}$	DESM-r$_{I\times O}$
case	.8	.7	.9	.9	.6	.9
opinion	.8	.7	.9	.9	.5	.8
paragraph	.6	.2	.5	.6	.5	.3
recursive	.4	.7	.6	.8	.7	1.0

Table 12: The table reports the mean values of λ_1 optimized during cross-validation. From these values it is clear that the sentence in its context has a lot of influence over the ranking decisions.

reject the null hypothesis (p-value=.898).

Overall, I was not able to show a statistically significant difference between the methods that operate on sentences only and the methods that take their contexts into account. However, this could also be due to the limited size of the data set. Otherwise, the results reported in Table 11 and Figure 25 do at least suggest the trend where taking sentences' context into account improves the performance. The influence of the context over the ranking decision was controlled by the hyperparameter λ_1, where $\lambda_1 = 1$ means that the ranking decision is based entirely on the sentence in its context whereas $\lambda_1 = 0$ means that the decision is based solely on the sentence itself. Table 12 reports the mean values of λ_1 as optimized during cross-validation. From these values it is clear that the sentence in its context had a lot of influence over the ranking decisions. This further corroborates the usefulness of the sentences' context in the task of retrieving sentences for statutory interpretation.

6.3 QUERY EXPANSION FOR NOVELTY DETECTION

In this section I investigate the effectiveness of focusing on the amount of information a sentence provides over what is known from the source provision. This closely ties to one of the rules in the definition of sentences' usefulness. Specifically, a sentence that does not

provide any additional information, over what is already known from the source provision, is considered to have 'no value.' I use several novelty detection techniques to establish if a sentence provides novel (i.e., additional) information with respect to the source provision. This means that for the purposes of these experiments I expand the query with the rest of the provision it comes from. In general, I am interested if and how could I improve the ranking by considering the novelty of a sentence. First, I investigate if down-ranking of 10% sentences with the lowest novelty scores improves the ranking. This appears appealing since it corresponds exactly with the rule from the definition. Second, I am also interested if sentences' novelty could be used as an indication of its value in general. This is somewhat less intuitive since it does not correspond directly to the definition of the usefulness. On the other hand, it is not completely absurd to think that the more novel information a sentence provides the higher the chance of it expressing something useful about the term of interest.

6.3.1 Experiments

As in the other sets of experiments I first retrieve all the sentences that contain the term of interest. I experiment with strategies to measure how much content the sentence adds with respect to the provision, as well as with measuring a distance between the two documents. As mentioned above the sentences' usefulness definition motivates the focus on novelty; sentences that do not provide additional information are deemed as having 'no value.' In [108] I observed that application of novelty detection techniques appeared to improve the performance of the ranking. However, an attempt to integrate a novelty detection-based ranker with a ranker focused on measuring similarity between a query and a sentence through linear interpolation failed. Interestingly, it was possible to successfully integrate the two when the novelty detection was used only to filter a small amount of the least valuable sentences. In [108] we did not investigate this issue any further—a gap that I fill with the new experiments presented here.

In [31] the author observes that applying novelty detection at the top positions of the rankings appears to be harmful. He proposes to use the results of the novelty detection only at the lower positions of the rankings. Here I would like to understand if this is indeed the

case and novelty detection mostly helps at the lower positions of the ranking and is less useful at the top positions. I use the below described novelty measures in two distinct ways:

1. Under the *filtering* condition (f) I place the 10% of the least novel sentences at the bottom of the ranking ordered from the most novel one. The remaining sentences are ordered randomly. As results I report an average of the 10,000 runs.
2. Under the *ranking* condition (r) I order all the sentences from the most novel to the least.

I acknowledge that the 10% threshold is chosen arbitrarily. I aim for such a threshold that encompasses a non-negligible amount of sentences, yet is sufficiently far from the top positions.

As the first approach to measure sentences' additional content, I evaluate the number of new words it includes, i.e., the number of words that appear in the sentence that are not contained in the provision. This simple measure is often surprisingly effective. [2]

$$NW = |\{w_i | w_i \in \mathbf{S}_j \setminus \{w_i | w_i \in \mathbf{P}\}\}|$$

$\mathbf{S}_j$ is the set of words w_i from a sentence and $\mathbf{P}$ is the set of words w_i from the source provision.

I also test a modified version of the above measure where I assign more weight to less common words as well as the words that appear closer to the query terms. The motivation is that less common words typically carry more information and the closer the word appears to a query term the more likely it may be suggestive about its meaning.

$$NWW = |\{w_i \cdot IDF \cdot 1/d | w_i \in \mathbf{S}_j \setminus \{w_i | w_i \in \mathbf{P}\}\}|$$

Here, the notation is the same as above; IDF is the inverse document frequency as defined above for the BM25 measure, and d is the distance (measured in words) from w_i to the closest query term (1 if they are neighbors).

The third approach is also closely related to the first one. The only difference is that I control for the size of the sentence. Thus, instead of the number of the new words I work with a ratio of the new words within a sentence.

$$NWR = \frac{NW}{|\{w_i | w_i \in \mathbf{S}_j\}|}$$

In addition, I measure Word Mover's Distance (WMD) on the embedding representations of sentence's words and words from a source provision. WMD captures the dissimilarity between two documents $\mathbf{d}$ and $\mathbf{d}'$ as the minimum distance that the embedded words of one document need to "travel" to reach the embedded words of another document. The distance is computed by solving the following linear program:

$$\begin{aligned} &\min_{\mathbf{T} \geq 0} \sum_{i,j=1}^{n} \mathbf{T}_{ij} c(i,j) \\ \text{subject to:} &\sum_{j=1}^{n} \mathbf{T}_{ij} = d_i \qquad \forall i \in \{1, \dots n\} \\ &\sum_{i=1}^{n} \mathbf{T}_{ij} = d'_i \qquad \forall j \in \{1, \dots n\} \end{aligned}$$

Here, $\mathbf{d}$ and $\mathbf{d}'$ are the bag of word vector representations of two documents (a sentence and the source provision in our case), $\mathbf{T} \in \mathbb{R}^{n \times n}$ is a flow matrix where $\mathbf{T}_{ij} \geq 0$ denotes how much of word i in $\mathbf{d}$ travels to word j in $\mathbf{d}'$. [62]

6.3.2 Results and Discussion

The results of the experiments described in Section 6.3.1 are reported in Table 13 (group and overall means) and Figure 26 (box and whisker plots and swarm plots). On the new larger data set I confirm the results from [108]. Overall, it appears that the models measuring the extra amount of words (count, weighted count, and ratio) perform better than WMD based models that measure the semantic distance between a sentence and the provision. Interestingly, almost all the methods except the WMD based ones outperform the Random by quite some margin (compare to the direct approach to sentence retrieval).

First, I examine the thesis statement S3.1 by testing the corresponding null hypothesis. As the first step I use the Friedman test to assess the hypothesis that at least one of the

Method	SmSp @10	SmSp @100	SmDs @10	SmDs @100	LgSp @10	LgSp @100	LgDs @10	LgDs @100	Overall @10	Overall @100
Random	.40 ± .07	.69 ± .15	.52 ± .08	.76 ± .11	.29 ± .16	.35 ± .18	.47 ± .11	.47 ± .11	.45 ± .12	.64 ± .20
NWrds-f	.49 ± .20	.73 ± .17	.55 ± .10	.79 ± .10	.30 ± .17	.36 ± .21	.51 ± .08	.52 ± .08	.49 ± .16	**.67 ± .21**
NWrdsW-f	.48 ± .19	.73 ± .16	.55 ± .09	.79 ± .10	.31 ± .18	.37 ± .20	.53 ± .07	.53 ± .08	.49 ± .15	**.67 ± .20**
NWrdsR-f	.45 ± .15	.71 ± .16	.53 ± .08	.77 ± .10	.30 ± .15	.36 ± .18	.49 ± .10	.50 ± .10	.47 ± .13	**.66 ± .20**
WMD-f	.40 ± .07	.69 ± .15	.52 ± .08	.76 ± .11	.29 ± .16	.35 ± .18	.46 ± .11	.47 ± .11	.45 ± .12	.64 ± .20
NWrds-r	.50 ± .22	.73 ± .19	.56 ± .18	.80 ± .11	.31 ± .21	.38 ± .22	.61 ± .16	.53 ± .10	.51 ± .20	.68 ± .21
NWrdsW-r	.53 ± .23	.72 ± .19	.56 ± .18	.80 ± .11	.32 ± .22	.36 ± .23	.64 ± .14	.54 ± .10	.53 ± .20	.68 ± .22
NWrdsR-r	.48 ± .27	.70 ± .22	.52 ± .19	.79 ± .11	.24 ± .21	.30 ± .21	.47 ± .29	.44 ± .22	.46 ± .24	.64 ± .25
WMD-r	.45 ± .18	.70 ± .18	.50 ± .21	.76 ± .11	.21 ± .21	.27 ± .21	.39 ± .28	.35 ± .23	.43 ± .22	.62 ± .25

Table 13: The table shows the results of the experiments on novelty detection. The NDGC@10 and NDGC@100 are shown for the small sparse queries (SmSp), small dense queries (SmDs), large sparse queries (LgSp), large dense queries (LgDs), and all of them together (Overall).

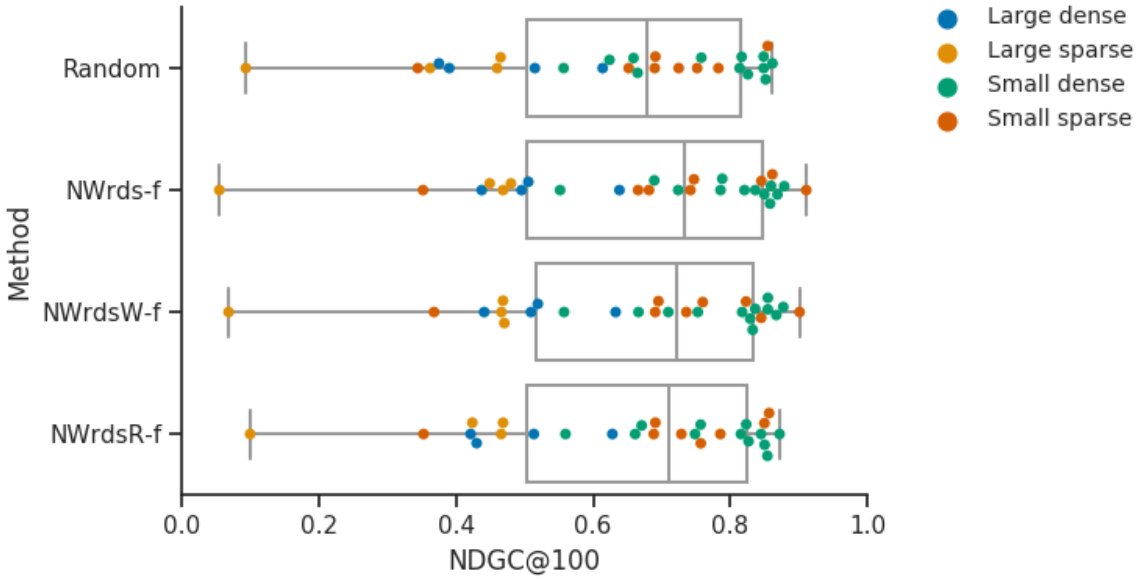

Figure 26: The figure shows scatter plots of the performance on the individual 28 queries measured in terms of NDGC@100. NWrds-f, NWrdsW-f, and NWrdsR-f appear to perform slightly better than the Random baseline.

Random, NWrds-f, NWrdsW-f, NWrdsR-f, WMD-f methods differs from the other four. The methods are compared in terms of their performance as measured with NDGC@100 on the Overall. I reject the null hypothesis (p-value=.0001) and conclude that there is at least one pair of methods that differ from each other. Since I am interested if any of the methods is different from the Random baseline I use the Holm procedure as described in Subsection 5.3. NWrdsW-f has the lowest p-value (.00004) when tested against the Random. When I adjust the p-value for 4 comparisons I reject the null for NWrdsW-f (p-value=.0001). Since I rejected the null hypothesis for NWrdsW-f I can step down to test NWrds-f, the method with the second lowest p-value (p-value=.0001). After applying the adjustment ($\alpha/(k-2)$) I reject the null for NWrds-f (.0003). I continue by stepping down to test NWrdsR-f (p-value=.006). After applying the adjustment ($\alpha/(k-3)$) I reject the null for NWrdsR-f (p-value=.01). Finally, I test the WMD-f method (p-value=.12) for which I cannot reject the null. Hence, I have established that the NWrds-f, NWrdsW-f, and NWrdsR-f are significantly different from Random.

Second, I analyze the thesis statement S4.2 by testing the corresponding null hypothesis. I will use the NWrdsW-f as the baseline since it tested as most likely different from Random. Again, as the first step I use the Friedman test to assess the hypothesis that at least one of the NWrdsW-f, NWrds-r, NWrdsW-r, NWrdsR-r, and WMD-r methods differs from the other ones. I reject the null hypothesis (p-value=.0000001) and conclude that there is at least one pair of methods that differ from each other. I use the Holm procedure to determine if any of the methods is different from the NWrdsW-f baseline. WMD-r has the lowest p-value (.000002) when tested against the NWrdsW-f. When I adjust the p-value for the 4 comparisons I reject the null for WMD-r (p-value=.00001). Since I rejected the null hypothesis for WMD-r I can step down to test NWrdsR-r, the method with the second lowest p-value (.008). After applying the adjustment ($\alpha/(k-2)$) I reject the null for NWrdsR-r (p-value=.024). I continue by stepping down to test NWrdsW-r (p-value=.34) for which I cannot reject the null. Hence, I have established that the WMD-r and NWrdsR-r are significantly different from NWrds-f (weaker performance).

My experiments indicate that models focused on "novel" sentences (i.e., those that provide additional information) perform better than the Random baseline. The models based

on novelty detection benefit from focusing on one aspect of the relevance definition and handle that aspect rather well. In the experiments on novelty the models measuring the extra amount of words performed better than models that measure the semantic distance between a sentence and the provision. This makes sense. The relevance definition focuses on the presence of additional information, which is related to but somewhat different from documents being distant in terms of their meaning.

I have confirmed that novelty detection methods produce better rankings than Random when used under the *filtering* condition, i.e., to down-rank the 10% least novel sentences. I have also shown that NWrdsR-r and WMD-r produce lesser quality ranking than NWrdsW-f. This does not mean that the methods perform better under the *filtering* condition as opposed to the *ranking* condition. This rather has to do with the fact that NWrds and NWrdsW appear to outperform NWrdsR and WMD irrespective of the condition. What is more interesting is that despite the slightly higher NDGC@100 statistics for the two better performing methods under the *ranking* condition (Table 13) I was not able to confirm they actually perform better than under the *filtering* condition. This would suggest that application of novelty detection is especially beneficial at lower positions of the rankings and has little to no effect at the higher positions. To investigate the phenomenon I measured the performance of all the four methods as I varied the *filtering* threshold from 0 to 100% (identical to *ranking* condition).

The results of the above described experiments are shown in Figure 27. The figure shows that it is the case that novelty detection benefits the ranking mostly by identifying less novel sentences and placing them at the bottom of the ranking. It appears there is no benefit in letting the novelty detection method inform the ranker as to which sentences should be placed at the top of the list. As I increased the *filtering* threshold I observed that the ranking performance improves until the threshold corresponds to approximately 50% (Figure 27). From then onward the improvement either appeared negligible (NWrds and NWrdsW) or the performance actually decreased (NWrdsR and WMD).

It appears that there is some fundamental difference between the workings of NWrds and NWrdsW methods and the workings of NWrdsR and WMD with respect to their effects on ranking at the top positions. The methods based on counting the new words (weighted or

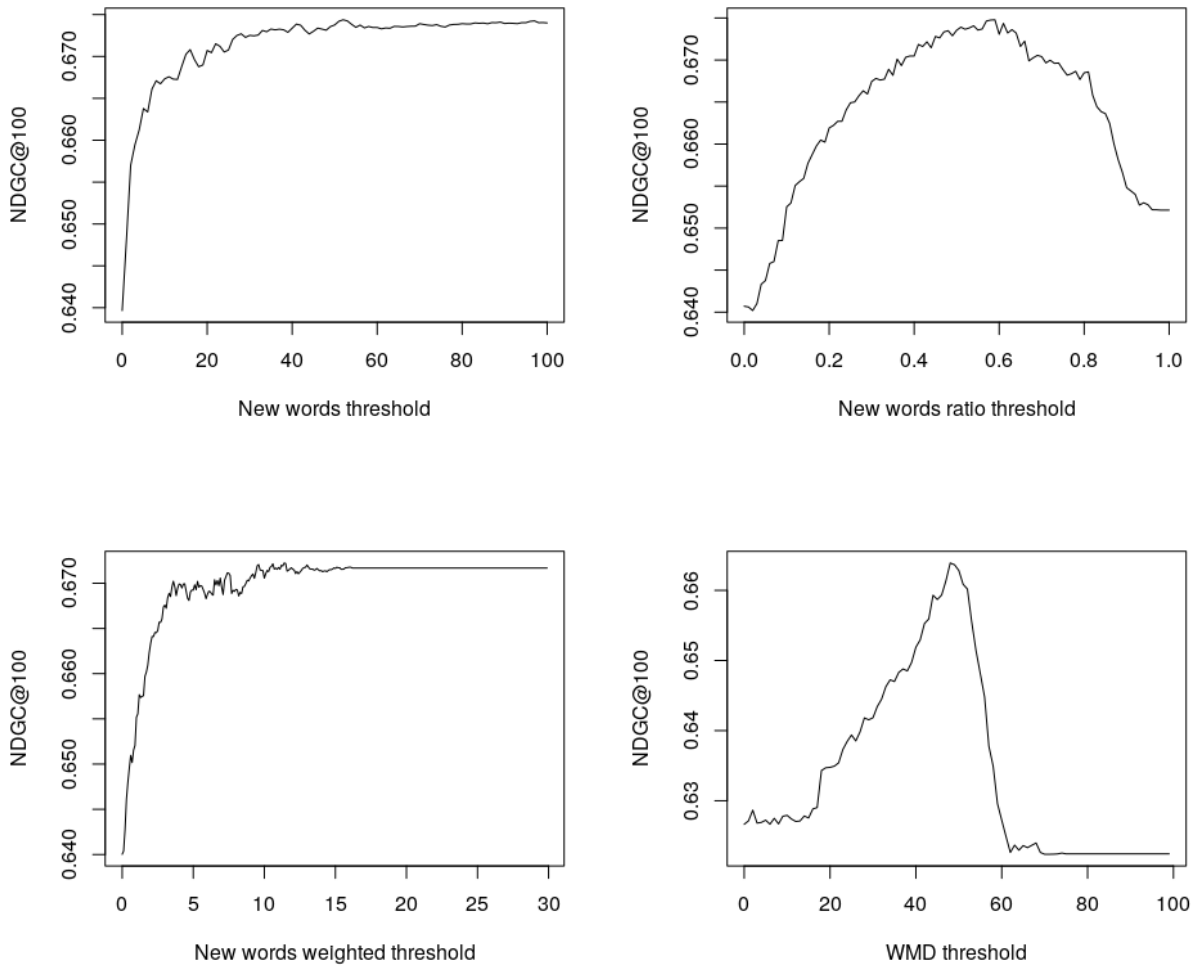

Figure 27: The figure shows performance in NDGC@100 for various levels of the novelty threshold as applied to the four novelty detection methods.

unweighted) appear to have comparable performance to Random. In case of WMD it appears that it would be more beneficial to switch the strategy at the top positions of the list and prefer sentences that are rather less distant from the query than the other way around. Intuitively this makes sense. For NWrdsR the results of the experiment appear to suggest that at the top positions it would be advantageous to prefer sentences with lower ratios of the

new words. One possible explanation could be that sentences where the ratios are too high come from the decisions where the term of interest is used in a different meaning (see the Subsections B.3.2 and B.3.3 in Appendix B and Section 6.4 for more details). However, it is not clear why the same phenomenon would not apply to the NWrds and NWrdsW methods.

From the qualitative standpoint, applying novelty detection leads to a better performance on a majority of the queries. The 'aural transfer' is one such query and it provides good insight into how do the methods benefit the rankings. The following examples illustrate the two classes of sentences that get detected as not useful (and that are consequently placed at the end of the ranking):

i. "Aural transfer."
[term of interest: "aural transfer;" gold label: 'no value']

ii. In Section 2510 "wire communication" is defined, in part, as "any aural transfer made in whole or in part through the use of facilities for the transmission of communications."
[term of interest: "aural transfer;" gold label: 'no value']

The first class (i) are very short sentences (e.g., headings) that by their very nature cannot provide any useful insight into the meaning of the term. Interestingly, these sentences would often top the rankings produced by the methods based on measuring similarity (Section 6.1). The second class (ii) are sentences that are verbatim citations or close paraphrases of the source provision. These sentences might sometimes be very tricky (such as the example shown above) because, apart from not providing any additional information, they would match the definition of a 'certain value' or 'high value' sentence. It is certainly a considerable improvement of the rankings once these types of sentences are prevented to appear at the high positions. Consequently, the queries that mostly benefited from the novelty detection methods are those with large numbers of the 'no value' sentences.

I established that novelty detection techniques are useful in identifying the least novel sentences which are almost always of little value for the interpretation. At the same time the techniques do not appear to be helpful for ranking sentences at the top positions. This confirms what we observed in [108] as well as what was reported in [31]. In addition, this corresponds to the definition of the sentences' usefulness for the interpretation of statutory terms.

6.4 QUERY EXPANSION FOR TOPIC SIMILARITY ASSESSMENT

In this section I investigate the effectiveness of extending the query with other words from the source provision on top of those that are part of the term of interest. In [108] using various similarity measuring strategies I ranked the sentences based on their similarity to the query as well as their similarity to the extended query. The motivating factor was the assumption that sentences that are "about" the term of interest from the provision are likely to contain other words from the provision or the words that are closely related to the term of interest. The assumption was shown to be wrong in the context of the task of retrieving case-law sentences for statutory interpretation. A linear interpolation was used to control how a model is informed by the original query versus the expanded query. When optimizing the hyperparameter λ_1 with respect to the development set the models always settled on ignoring the expanded query altogether. This made the models equal to those that retrieve the sentences directly. [108]

Instead of comparing the similarities of the query and the expanded query to the sentences themselves, I showed in [108] that it is more interesting to compare them to the whole cases. The intuition behind this approach was that, given I only work with sentences that mention the term of interest, the cases that are most similar to the expanded queries are those that focus on the source provision. [108] These are almost guaranteed to use the term of interest in the same meaning as the source provision. And this corresponds to part of the definition of the sentences' usefulness. Specifically, a sentence that mentions the term of interest in a different yet related meaning should be demoted to a category one step down, while a sentence that mentions the term of interest in a completely different meaning has 'no value' (see Appendix B for details).

In general, I am interested if and how could I improve the ranking by considering the topic similarity between the source provision and the full text of a decision the sentence comes from. First, I investigate if down-ranking of sentences that come from a decision the topic similarity of which is less than 50% of the most similar ones—the approach I used in [108] (see below for more details). This appears appealing since it corresponds closely with the rule from the definition. Second, I am also interested if the full-text decisions'

topic similarity could be used as an indication of its value in general. This is somewhat less intuitive since it does not correspond directly to the definition of the usefulness. On the other hand, it is not completely absurd to think that the more similar the decision is in terms of its topics to the source provision the higher the chance of it containing the sentence that expresses something useful about the term of interest.

6.4.1 Experiments

As in case of the preceding sets of experiments, I first retrieve all the sentences that contain the term of interest. To measure the similarity I will use some of the models presented in the section on ranking the sentences directly. Specifically, I will use the BM25 and the TF-ISF methods from those that match the query terms only, and GloVe, fastt^{SIF}, DESM$_{I\times I}$, and DESM$_{I\times O}$ that compare all the words from the document. The selection of the methods was done with the aim to cover the methods that can be expected to behave differently while avoiding unnecessary duplication of methods that would most likely produce similar results. The inclusion of the TF-ISF may seem counter-intuitive since this measure is specifically tailored for matching short documents which is not the case in this set of experiments. The reason why I include this method is that it performed the best in similar experiments in [108]. The specific implementations of the models measuring the similarities between expanded queries and the whole cases are the same as those presented in Subsection 6.1.1 (direct ranking). The only difference is that the query (originally the term of interest) is replaced with the expanded query and sentences are replaced with full texts of the cases.

There is a number of techniques focused on modeling abstract topics that appear in a document collections. Some well-established methods are Latent Semantic Analysis (LSA) [21], its probabilistic version (pLSA) [48], and Latent Dirichlet Allocation (LDA) [8]. The goal of topic modelling is to discover compact representations of documents while preserving the essential statistical relationships useful for tasks such as classification, novelty detection, summarization, and similarity and relevance judgments. [8] A document is then represented as a projection into a fixed-size topic space as opposed to a projection into a word occurrence (count, possibly weighted) space. The main difference is that the size of the topic space

(typically several hundred dimensions) is many times smaller than the word space (the number of dimensions corresponds to the number of unique tokens in the corpus). In this respect the topic models are similar to the word embedding models. The main difference is that while topic models focus on the whole documents the word embedding models use rather small windows (narrow contexts) to discover word collocations. Whereas the primary goal of the topic models is to model the relationships among documents the primary goal of word embeddings is to model the relationships among words.

LSA was often criticized for its lack of probabilistic interpretation and, hence, pLSA was developed as a response. pLSA models each word in a document as a sample from a mixture of multinomial random variables ("topics"). While this representation models the probabilities at the word level it is incomplete because it does not allow for probabilistic interpretation at the level of documents. LDA is a generative probabilistic model of a corpus. Documents are represented as random mixtures over latent topics. A topic is understood as a distribution over words. [8] Hence, in my experiments on topic similarity assessment I analyze performance of LDA as quite an appealing method to model the phenomenon.

To facilitate the experiments I used the LDA model implementation from gensim[7] [98] based on [47]. I trained the model with 300 topic dimensions in a single pass over the 6,715,418 full case text documents from the data set described in Chapter 4. An example of 5 most important topics with 5 most important words is presented in Table 14. It is encouraging that the topics appear to be quite sensible and a lawyer could easily recognize why each topic is being included and what it likely represents.

6.4.2 Results and Discussion

The results of the experiments described in Section 6.4.1 are reported in Table 15 (group and overall means) and Figure 15 (box and whisker plots and swarm plots). On the new larger data set I confirm the results from [108] showing that matching the expanded queries to the whole cases proves to be a viable strategy. From the raw NDGC numbers it appears that the methods are competitive with those retrieving the sentences directly (Section 6.1)

[7] https://radimrehurek.com/gensim/models/ldamodel.html

argumentation	*trusts*	*unions*	*judges*	*commerce*
defendant	trust	union	judge	trade
deny	trustee	labor	justice	customer
contention	create	strike	magistrate	price
upon	principal	national	judicial	market
court	power	america	chief	consumer

Table 14: The table shows an example of 5 most important topics with 5 most important words from the LDA model trained on the full case texts.

or those focused on novelty (Section 6.3). This is surprising since these methods do not even operate on the level of individual sentences and each sentence coming from a single case gets assigned the same score (such sentences are then ranked per the order of their appearance). In [108] we showed that these models have an interesting property of preferring cases that discuss the source provision or the term of interest as opposed to the cases that mention a term identical with the term of interest that comes from a different domain. This appears to be a serious problem the treatment of which is enough to make these methods perform similarly to those that really attempt to rank the sentences themselves.

First, I examine the thesis statement S4.1 by testing the corresponding null hypothesis. As the first step I use the Friedman test to assess the hypothesis that at least one of the Random, BM25-f, TF-ISF-f, GloVe-f, fastt-f^{SIF}, DESM-f$_{I\times I}$, DESM-f$_{I\times O}$, and LDA-f methods differs from the other seven. The methods are compared in terms of their performance as measured with NDGC@100 on the Overall. I reject the null hypothesis (p-value=.019) and conclude that there is at least one pair of methods that differ from each other. Since I am interested if any of the methods is different from the Random baseline I use the Holm procedure as described in Subsection 5.3. The method with the lowest p-value (.0016) is DESM-f$_{I\times O}$. When I adjust the p-value for 7 comparisons I reject the null for DESM-f$_{I\times O}$ (p-value=.011). Since I rejected the null hypothesis for DESM-f$_{I\times O}$ I can step down to test LDA-f, the method with the second lowest p-value (.0019). After applying the adjustment

Method	SmSp @10	SmSp @100	SmDs @10	SmDs @100	LgSp @10	LgSp @100	LgDs @10	LgDs @100	Overall @10	Overall @100
Random	.40 ± .07	.69 ± .15	.52 ± .08	.76 ± .11	.29 ± .16	.35 ± .18	.47 ± .11	.47 ± .11	.45 ± .12	.64 ± .20
BM25-f	.40 ± .14	.70 ± .15	.50 ± .08	.75 ± .12	.37 ± .06	.44 ± .06	.46 ± .18	.46 ± .18	.45 ± .12	.65 ± .18
TF-ISF-f	.41 ± .12	.72 ± .14	.49 ± .08	.74 ± .13	.34 ± .11	.41 ± .10	.45 ± .21	.44 ± .21	.44 ± .12	.64 ± .20
GloVe-f	.39 ± .08	.69 ± .15	.53 ± .08	.78 ± .10	.30 ± .15	.35 ± .17	.47 ± .11	.48 ± .12	.45 ± .12	.65 ± .20
fastt-f^{SIF}	.41 ± .07	.70 ± .16	.51 ± .08	.77 ± .11	.32 ± .14	.38 ± .15	.49 ± .15	.50 ± .14	.45 ± .12	**.65 ± .19**
DESM-f$_{I\times I}$	.39 ± .10	.68 ± .16	.51 ± .08	.76 ± .11	.30 ± .15	.37 ± .16	.49 ± .11	.48 ± .11	.44 ± .12	.64 ± .19
DESM-f$_{I\times O}$	.38 ± .13	.69 ± .16	.52 ± .08	.77 ± .11	.32 ± .12	.39 ± .13	.51 ± .12	.51 ± .11	.45 ± .13	**.66 ± .18**
LDA-f	.43 ± .05	.73 ± .11	.52 ± .07	.77 ± .11	.36 ± .08	.44 ± .06	.50 ± .10	.51 ± .11	.47 ± .09	**.67 ± .16**
BM25-r	.45 ± .25	.72 ± .19	.50 ± .13	.74 ± .13	.31 ± .21	.48 ± .06	.45 ± .20	.45 ± .26	.45 ± .19	.66 ± .20
TF-ISF-r	.41 ± .13	.72 ± .14	.46 ± .12	.73 ± .14	.24 ± .15	.42 ± .11	.40 ± .25	.46 ± .23	.41 ± .16	.64 ± .19
GloVe-r	.35 ± .19	.69 ± .17	.48 ± .11	.76 ± .11	.38 ± .08	.52 ± .03	.45 ± .18	.52 ± .23	.42 ± .15	.67 ± .17
fastt-r^{SIF}	.36 ± .13	.68 ± .18	.49 ± .11	.77 ± .11	.44 ± .10	.51 ± .05	.53 ± .34	.54 ± .23	.45 ± .17	.67 ± .17
DESM-r$_{I\times I}$	.35 ± .15	.68 ± .15	.46 ± .11	.76 ± .12	.40 ± .18	.49 ± .05	.50 ± .33	.53 ± .21	.43 ± .17	.66 ± .17
DESM-r$_{I\times O}$	.39 ± .21	.69 ± .16	.60 ± .14	.80 ± .12	.38 ± .06	.48 ± .03	.54 ± .28	.57 ± .16	.50 ± .20	.69 ± .17
LDA-r	.36 ± .12	.71 ± .14	.48 ± .10	.76 ± .11	.29 ± .11	.41 ± .07	.48 ± .20	.53 ± .15	.42 ± .14	.66 ± .17

Table 15: The table shows the results of the experiments on topic similarity. The NDGC@10 and NDGC@100 are shown for the small sparse queries (SmSp), small dense queries (SmDs), large sparse queries (LgSp), large dense queries (LgDs), and all of them together (Overall).

($\alpha/(k-2)$) I reject the null for LDA-f (.011). I continue by stepping down to test fastt-f^{SIF} (p-value=.007). After applying the adjustment ($\alpha/(k-3)$) I reject the null for fastt-f^{SIF} (.035). I continue by stepping down to test GloVe-f (p-value=.06) for which I cannot reject the null. Hence, I have established that DESM-f$_{I\times O}$, LDA-f, and fastt-f^{SIF} are significantly different from Random.

Second, I analyze the thesis statement S4.2 by testing the corresponding null hypothesis. I will use the DESM-f$_{I\times O}$ as the baseline since it tested as most likely different from Random. Again, as the first step I use the Friedman test to assess the hypothesis that at least one of the DESM-f$_{I\times O}$, BM25-r, TF-ISF-r, GloVe-r, fastt-r^{SIF}, DESM-r$_{I\times I}$, DESM-r$_{I\times O}$, and LDA-r methods differs from the other ones. Interestingly, we cannot reject the null hypothesis (p-value=.415). This is somewhat surprising since most of the methods appear to perform much better under the *ranking* as opposed to the *filtering* condition. Both, the mean NDGC@100 scores (Table 15) as well as the swarm plots (Figure 28), confirm this. However, a closer inspection of the swarm plots enables one to understand what is happening. This is because there is a small number of large sparse queries, especially the 'identifying

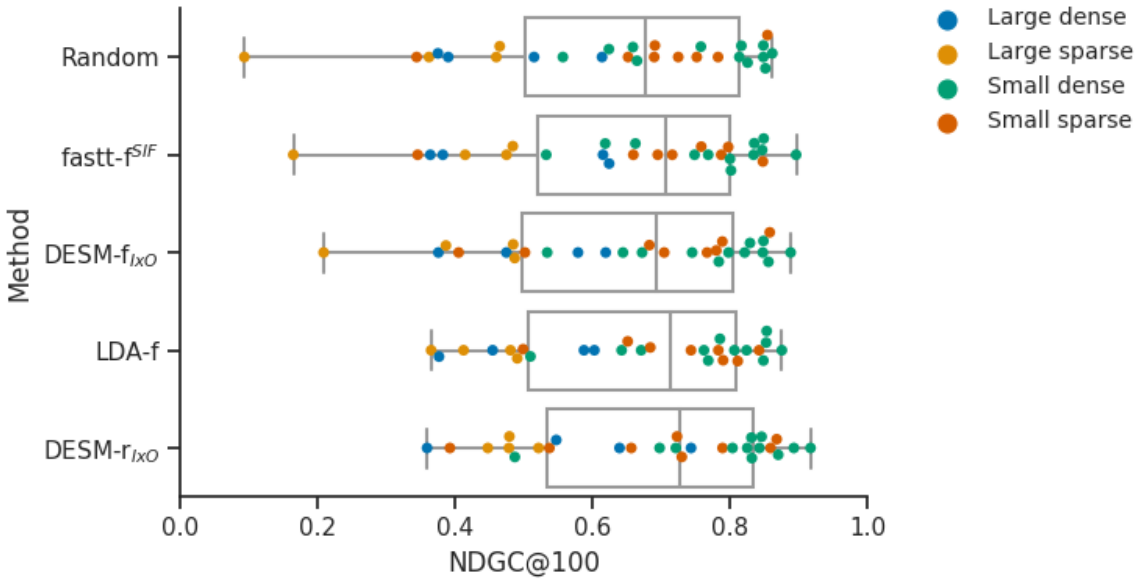

Figure 28: The figure shows scatter plots of the performance on the individual 28 queries measured in terms of NDGC@100. fastt-f^{SIF}, DESM-f$_{I\times O}$, and LDA-f appear to perform slightly better than the Random baseline. DESM-r$_{I\times O}$ appears to perform even better (not significant).

particular,' that appear to be really difficult (the orange point below 0.2 as shown in Figures 24 and 26). These queries obviously suffer from many sentences where the term is used in a different context because the performance on these queries is much better under most of the ranking methods studied in this section. The improvement in NDGC@100 scores for these queries is so high that it accounts for .02–.03 improvement in the Overall NDGC@100. As for the rest of the queries the gain in performance does not appear to be that radical. Quite contrary, if one looks at the large swarm of small dense queries at the top of the Random swarm plot in Figure 28 and compares it to the corresponding groups in the methods tested under the ranking condition, he or she would observe that the performance of these appears to be slightly higher under Random.

The Overall NDGC@100 scores reported in Table 15 suggest that methods matching the query (expanded with the source provision) to full case texts could be used not only for

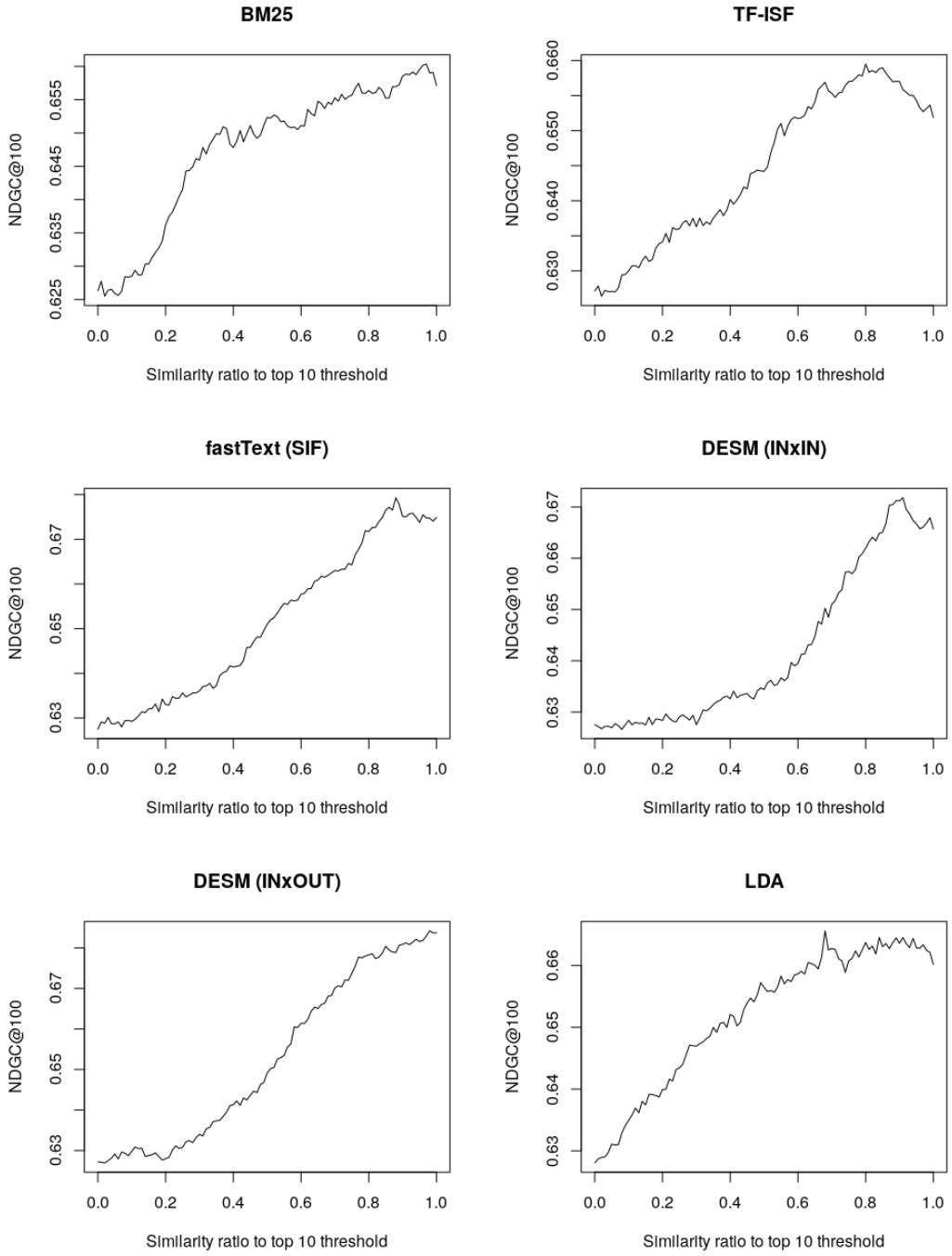

Figure 29: The figure shows performance in NDGC@100 for various levels of the topic similarity threshold as applied to the six topic similarity measurement methods.

filtering but as full-fledged ranking procedures as well. This definitely appears to be the case with the methods that consider all the words in the full text documents (i.e., the word embedding based ones and LDA). The two methods that consider the query terms only (BM25 and TF-ISF) appear to bring the most value by filtering out the sentences that are obviously not useful (i.e., those that likely mention the term of interest in different sense). When applied to the whole sentence list there is only minimal further improvement. However, I was not able to show that any of the *ranking* methods (-r) significantly outperforms DESM-f$_{I\times O}$ (one of the best *filtering* methods). Further investigations revealed that the methods most likely do not work too well as rankers. The apparent increase in the Overall NDGC@100 scores was shown to be attributable to a dramatic increase in scores for just a few terms of interest as opposed to them working better in general. To gain additional insight into the phenomenon I measured the performance of all the four methods as I varied the *filtering* threshold from 0 to 100% (almost identical to ranking condition).

The results of the above described experiments are shown in Figure 29. The figure shows that it is mostly the case that the performance increases as the threshold shifts higher. It is also apparent that in case of the BM25, TF-ISF, and LDA methods the increase plateaus when the threshold reaches 50%. This is not the case with other methods. However, what is universal across all the methods, except the DESM$_{I\times O}$, is that there is a visible dip in performance as the threshold approaches 100%. This clearly shows that application of the methods at the very top of the rankings harms the performance. Even the DESM$_{I\times O}$ visibly plateaus around 80%. The reason there is no performance dip towards the end for this method could be its exceptional improvement at the small number of the lowest performing queries (e.g., 'identifying particular').

From the qualitative standpoint, applying topic similarity assessment leads to a better performance on a number of queries. The ones that benefit the most are those that consist of more general terms that can easily attach to many different meanings (e.g., "preexisting work," "electronic signature"). The following examples illustrate the kind of sentences that are correctly recognized as using the term in a completely different meaning:

i. In that regard, Dr. Vale opined that the combination of Employee's occupational disease of September 2, 1994, and her preexisting work and nonwork–related conditions rendered her totally disabled. [term of interest: "preexisting work;" gold label: 'no value']

ii. During this meeting the Lockheed Martin and Gibbs & Cox representatives emphasized key features of their proposed littoral combat ship, including low-speed stability features, sprint speeds of 40 to 50 knots over 1,000 nautical miles, "[l]ow electronic signature," and price.
[term of interest: "electronic signature;" gold label: 'no value']

In the first case the mentioned term comes from the area of Employment Law while the term of interest is from the area of Copyright. The meanings are completely different. In the second example the term of interest refers to the "electronic signature" that is used to sign documents while the mentioned term relates to the "electronic signature" of a ship. It is certainly a considerable improvement of the rankings once these types of sentences are prevented to appear at the high positions. Consequently, the queries that mostly benefited from the topic similarity assessment methods are those with large number of low value sentences that mention the term of interest in a different meaning.

I established that topic similarity assessment methods are useful in identifying the sentences that mention the term of interest in a different meaning. At the same time the techniques do not appear to be helpful for ranking sentences at the top positions. This confirms what we observed in [108]. In addition, this corresponds to the intuition about the definition of the sentences' usefulness for the interpretation of statutory terms.

6.5 USING FUNCTIONAL SEGMENTATION OF COURTS DECISIONS FOR SENTENCE FILTERING

Court opinions consist of several high-level parts each of which has a different function. Distinguishing the functional parts is crucial for a lawyer to be able to focus attention on the pieces of the opinion that matter. Segmentation of case texts into functional parts could be a useful component in the sentence discovery pipeline. Mainly, this is because the segments provide clues to the meaning of their contents [6, 42]. For instance, knowing in which high-level part a sentence appears would help to annotate sentences in terms of the roles they play in legal argument. Sentences that state a finding of fact or state a legal rule are more likely to be found in the Background or Analysis parts, respectively. We introduced the task and processing pipeline for segmenting U.S. court decisions into functional and issue specific

parts in [105]. Here, I use part of the pipeline focused on recognizing the functional parts. I plan to employ the results of the segmentation in the sentence ranking pipeline. However, this line of work has much wider applicability. It has been recognized as "important but often neglected" in [110] where the author speaks about segmentation of a legal document into its structural elements such as, e.g., facts, arguments, and rulings. As explained in [68], "The ability to identify and partition a document into segments is important for many NLP tasks, including IR (this work), summarization, and text understanding."

The goal of the segmentation is to split the text of an opinion into a varying number of consecutive non-overlapping parts. Each part is assigned one of the following types [105]:

1. **Introduction** – the opening part which typically consists of lines indicating the deciding court, judges, the case citation, parties, etc. It is not uncommon that the court would include a summary of the decision.
2. **Background** – the part where the court describes the procedural history of the case, the relevant facts, as well as what the parties are claiming. Its tone is usually descriptive, i.e., the court refrains from expressing its own opinions.
3. **Analysis** – the part where the court discusses and reasons about the issues of the case and states its outcome. Quite often the tone is deliberative, i.e., the court expresses opinions on the issues, arguments, or claims. The court may deal with a single issue or it may treat several issues separately.
4. **Concurrence or Dissent** – the part where opinions of concurring or dissenting judges are presented. There may be dedicated sections for each concurrence/dissent but there could be just a single sentence informing about the list of concurring/dissenting judges.
5. **Footnotes** – a list the indices of which are references to different parts of a decision. Each item of the list provides additional information as to what is in the text at the place of reference.
6. **Appendix** – a separate document attached to a decision as supplemental material.

In this section I am interested if the information about the functional parts sentences come from (see [105] for details) could be useful in their ranking. Note that this approach does not assign a sentence with a score. Hence, it cannot be used to rank the sentences

directly. However, I expect that while some of the parts could be a very likely source of useful sentences (e.g., the Analysis part), other parts much less so (e.g., the factual and procedural Background). Consequently, one could at least use the information to down-rank sentences coming from the parts that are unlikely sources of useful sentences.

6.5.1 Experiments

For the experiments I use the segmentation pipeline developed in [105]. I used the described procedure to segment each decision into its Introduction, Background, and Analysis parts. Furthermore, I use the manually annotated segments created by the Caselaw access project team.[8] Specifically, I use their segmentation of the decisions into the individual opinions (e.g., majority opinion, concurrence, dissent). Then, each sentence is labeled with respect to the functional part it comes from. I use dedicated categories for Introduction, Background, and Analysis. For the remaining parts I use a catch-all Others category. The distribution of all the 26,959 sentences according to the part they come from is shown in Figure 30. Apparently, a large majority of sentences comes from the main Analysis part. This appears reasonable.

Altogether, I experiment with four methods where each corresponds to one of the categories. For each of the methods sentences from a specific category are first separated from the rest. This results in two sentence lists for each of the queries. Within a query both lists are randomly shuffled and then concatenated in a way where the list containing the sentences from the specified category comes on top. I repeat the procedure for 100 times and report the mean NDGC scores as results. The idea behind this procedure is that if a specific functional part contains on average much more or much less valuable sentences this should show up in an increased or decreased NDGC scores.

6.5.2 Results and Discussion

The results of experiments described in Section 6.5.1 are reported in Table 16. I examine the thesis statement S5. by testing the corresponding null hypothesis. As the first step I use

[8] case.law

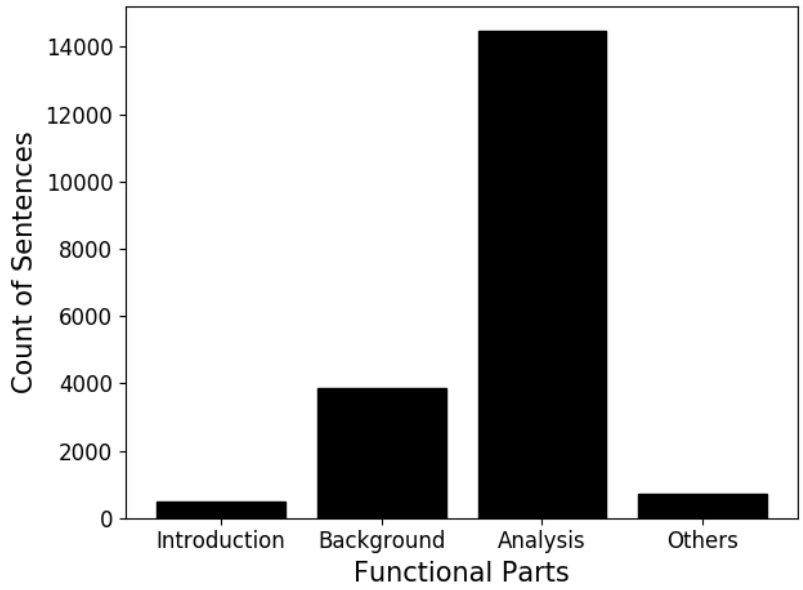

Figure 30: The figure shows distribution of all the 26,959 sentences according to the functional part they come from.

the Friedman test to assess the hypothesis that at least one of the Random, Introduction, Background, Analysis, and Others methods differs from the other five. The methods are compared in terms of their performance as measured with NDGC@100 on the Overall. As expected I cannot reject the null hypothesis (p-value=.154) and conclude that all the methods appear to perform similarly. More importantly, I cannot show that any of the methods perform differently from Random.

The above presented conclusion is somewhat disappointing. However, the distribution of sentence values with respect to the functional parts (Figure 31) reveals some interesting, perhaps even surprising, insights. First, obviously it is possible for a 'high value' sentences to appear in the Introductory and Background parts of the decision. Second, the majority of 'no value' sentences does appear in the Analysis part. Although, I cannot use the information about the functional part a sentence comes from as an effective filtering criterion, it might still prove to be useful as a feature in a learning-to-rank system (see Chapter 7).

Method	SmSp		SmDs		LgSp		LgDs		Overall	
	@10	@100	@10	@100	@10	@100	@10	@100	@10	@100
Random	.40 ± .07	.69 ± .15	.52 ± .08	.76 ± .11	.29 ± .16	.35 ± .18	.47 ± .11	.47 ± .11	.45 ± .12	.64 ± .20
Introduction	.39 ± .06	.68 ± .15	.49 ± .10	.75 ± .10	.28 ± .15	.34 ± .17	.47 ± .12	.48 ± .10	.43 ± .12	.63 ± .19
Background	.40 ± .07	.68 ± .17	.51 ± .11	.77 ± .09	.29 ± .18	.33 ± .20	.47 ± .13	.48 ± .12	.44 ± .14	.64 ± .21
Analysis	.40 ± .07	.69 ± .15	.53 ± .10	.77 ± .12	.30 ± .16	.35 ± .17	.46 ± .12	.46 ± .12	.45 ± .13	.64 ± .20
Others	.43 ± .12	.69 ± .15	.51 ± .13	.76 ± .12	.28 ± .18	.33 ± .21	.41 ± .20	.43 ± .20	.44 ± .16	.63 ± .22

Table 16: The table shows the results of the experiments on ranking with functional segmentation. The NDGC@10 and NDGC@100 are shown for the small sparse queries (SmSp), small dense queries (SmDs), large sparse queries (LgSp), large dense queries (LgDs), and all of them together (Overall).

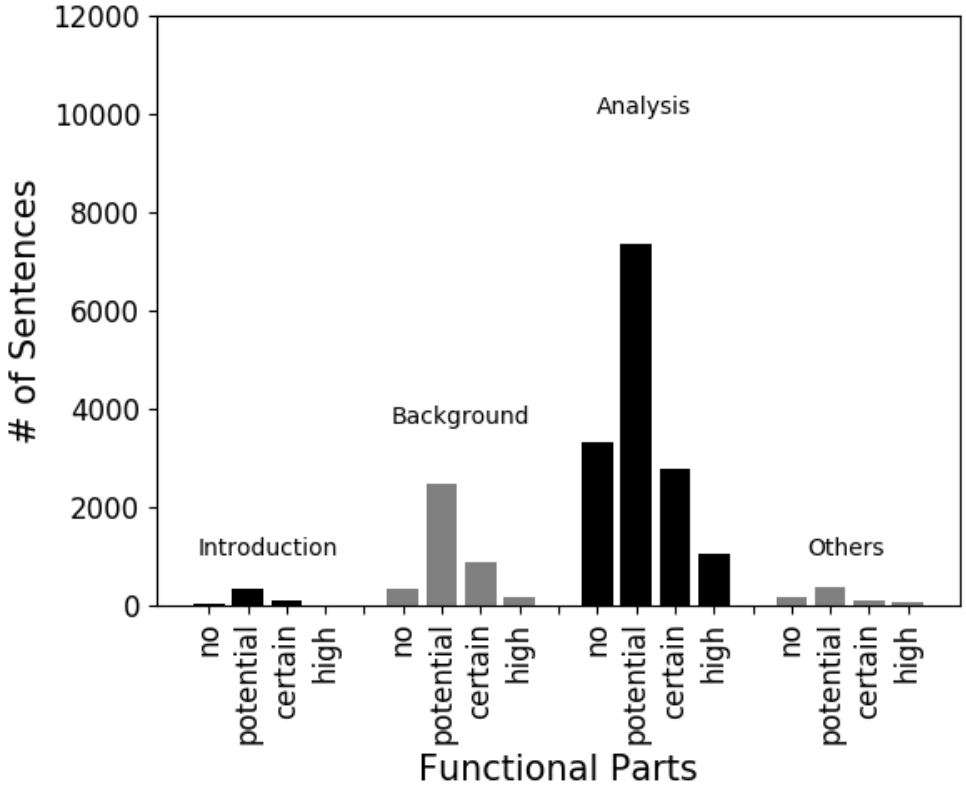

Figure 31: The figure shows sentence value distribution over the functional parts.

6.6 COMPOUND MODELS

In the final batch of experiments I explore the effects of combining models evaluated previously. In [74] it was shown that there is a notable difference between an identical ranker being applied to a full document set versus being applied only at the most relevant documents as deemed by some other ranker (telescoped data set). In [79] it was shown that a ranking method based on measuring similarity between the centroids of the embedding vectors of a query and a document performed reasonably well when applied at the top positions of the result list retrieved by an established ranking algorithm. The very same method performed poorly when applied on the full list. In this work the document lists associated with each query are somewhat telescoped because they only consist of sentences that contain at least one full match of the respective query. At the same time no ranking method has really been applied to the lists and hence these would not qualify as telescoped in the sense of [74] and [79]. In my experiments on using novelty detection and topic similarity assessment I have established that the methods are useful for filtering out sentences that are not valuable for the interpretation of statutory terms. At the same time I have observed that the methods are somewhat problematic when used as rankers, i.e., to make decision about the ranking of sentences at the top of the lists. Hence, it appears reasonable to use these methods to telescope the results lists and improve the performance of the base ranker such as one of the context-aware methods presented in Section 6.2.

6.6.1 Experiments

For each sentence I compute scores based on the most successful models that (1) utilize the sentence's context (Section 6.2), (2) measure the sentence's novelty with respect to the source provision (Section 6.3), (3) measure the topic similarity between the source provision and the full case text. Note that the same strategy was used in [108]. There the motivation was to inform the relatively successful context-aware models with different types of signals coming from the models based on novelty detection and topic similarity assessment. Here, I was not able to reliably establish that the context-aware methods do outperform the methods that

consider the sentences only (see Section 6.2.2 for details). Despite this the context-aware methods do appear to perform better and hence, I use them as we din in [108]. Based on the error analysis performed in [108] we proposed a task specific sentence retrieval model that has the following general form:

$$\text{CMP-rank}(q, s, sp, \mathbf{C}) = \textit{Sim-i}(q, s, \mathbf{C}_i) \cdot NI(s, sp) \cdot DDI(q, \mathbf{C}_j)$$

Here, q is the query, s is a sentence, sp is the source provision, $\mathbf{C}$ is the set of available contexts ($\mathbf{C}_i$ is a specific context where $i \in \{c, o, p, r\}$), *Sim-i* is the base ranking function from the family of the context-aware models, NI is the novelty indicator function from the family of the novelty detection models, and DDI is the different domain indicator function from the group of the models utilizing query expansion. NI mitigates the problem of retrieving sentences that are not much more than partial or complete citations of the source provision. DDI tackles the problem of retrieving cases that discuss the term that is the same as the term of interest, yet it comes from a different domain (see Sections 6.3.2 and 6.4.2 for discussion on these problems).

As concrete implementations of *Sim-i* on which to experiment I chose BM25-p and GloVe-R which are the most successful models from their respective categories (only query terms matching versus full document matching).

I utilize NWrdsW-f for the NI component. This gives the following implementation of NI:

$$NI(s, sp) = \begin{cases} 1 & \text{if } NWrdsW - f(s, sp) \geq \lambda_1 \\ 0 & \text{if } NWrdsW - f(s, sp) < \lambda_1 \end{cases}$$

This component requires a sentence to surpass a set novelty threshold as measured by NWrdsW-f. The threshold is controlled through hyperparameter λ_1. I use the same threshold (10 words) as in the filtering experiment (Section 6.3.1). Note that this threshold is rather conservative as could be observed from Figure 27. I opt for the conservative measure as to stay on the safe side with respect to the conclusions I would draw from this experiment. If the sentence fails to surpass the threshold it is moved towards the end of the results list.

The "discarded" sentences are among themselves ordered by the score assigned by the base ranker.

I use LDA-f as the DDI component. This yields the following implementation of DDI:

$$DDI(s,c) = \begin{cases} 1 & \text{if LDA-f(sp,c)} \geq \lambda_2 Avg_{10} \\ 0 & \text{if LDA-f(sp,c)} < \lambda_2 Avg_{10} \end{cases}$$

Here, Avg_{10} is the average of the scores of the top 10% documents. This approximates the score of a document that comes from the same domain as the term I am interested in. The reason I do not take the score of the top document is to account for the possibilities of a case citing the source provision multiple times or of a short case the majority of which is the citation of the source provision. These would give an unrealistically high estimate. The hyperparameter λ_2 then controls the threshold below which I consider documents to come from a different domain. Based on the error analysis performed in [108] I set $\lambda_2 = 0.5$ for the experiments. As in case of the λ_1 this is a relatively conservative choice (see Figure 29).

6.6.2 Results and Discussion

The results of the experiments described in Section 6.6.1 are reported in Table 17 and Figure 32 (box and whisker plots and swarm plots). On the new larger data set I confirm the results from [108]. The new results indicate that it is indeed the case that augmenting the base context-aware methods with the approaches focused on novelty and topic similarity leads to models that are stronger than any of their constituent parts.

First, I examine the thesis statement S6.1 by testing the corresponding null hypothesis. As the first step I use the Friedman test to assess the hypothesis that at least one of the BM25-p, NWrdsW-f, LDA-f, and BMp+NW+LDA methods differs from the other three. The methods are compared in terms of their performance as measured with NDGC@100 on the Overall. I reject the null hypothesis (p-value=.001) and conclude that there is at least one pair of methods that differ from each other. Since I am interested if BMp+NW+LDA is different from any of the three baselines I use the Holm procedure as described in Subsection 5.3. LDA-f has the lowest p-value (.0004) when tested against the BMp+NW+LDA. When

Method	SmSp @10	SmSp @100	SmDs @10	SmDs @100	LgSp @10	LgSp @100	LgDs @10	LgDs @100	Overall @10	Overall @100
BM25-p	.51 ± .12	.74 ± .13	.64 ± .14	.81 ± .11	.59 ± .16	.54 ± .05	.64 ± .27	.59 ± .18	.60 ± .16	.72 ± .16
GloVe-r	.42 ± .15	.75 ± .09	.56 ± .07	.79 ± .10	.58 ± .23	.59 ± .09	.45 ± .35	.50 ± .25	.51 ± .18	.71 ± .16
NWrdsW-f	.48 ± .19	.73 ± .16	.55 ± .09	.79 ± .10	.31 ± .18	.37 ± .20	.53 ± .07	.53 ± .08	.49 ± .15	.67 ± .20
LDA-f	.43 ± .05	.73 ± .11	.52 ± .07	.77 ± .11	.36 ± .08	.44 ± .06	.50 ± .10	.51 ± .11	.47 ± .09	.67 ± .16
BMp+NW	.54 ± .22	.75 ± .15	.63 ± .14	.81 ± .11	.55 ± .09	.52 ± .08	.63 ± .25	.59 ± .11	.59 ± .17	.72 ± .16
GVr+NW	.50 ± .19	.75 ± .11	.61 ± .12	.80 ± .10	.60 ± .15	.57 ± .03	.52 ± .33	.57 ± .19	.57 ± .18	.72 ± .15
BMp+LDA	.54 ± .15	.78 ± .09	.61 ± .15	.80 ± .12	.59 ± .18	.56 ± .06	.68 ± .11	.61 ± .09	.60 ± .15	.73 ± .14
GVr+LDA	.48 ± .16	.75 ± .11	.53 ± .06	.78 ± .10	.56 ± .25	.58 ± .10	.55 ± .17	.58 ± .12	.52 ± .14	.72 ± .13
BMp+NW+LDA	.55 ± .11	.78 ± .12	.64 ± .14	.82 ± .10	.58 ± .16	.56 ± .02	.65 ± .23	.62 ± .11	.61 ± .15	.74 ± .14
GVr+NW+LDA	.45 ± .15	.77 ± .09	.54 ± .07	.79 ± .09	.57 ± .24	.60 ± .09	.45 ± .32	.53 ± .19	.50 ± .17	.72 ± .14

Table 17: The table shows the results of the experiments with compound models. The NDGC@10 and NDGC@100 are shown for the small sparse queries (SmSp), small dense queries (SmDs), large sparse queries (LgSp), large dense queries (LgDs), and all of them together (Overall).

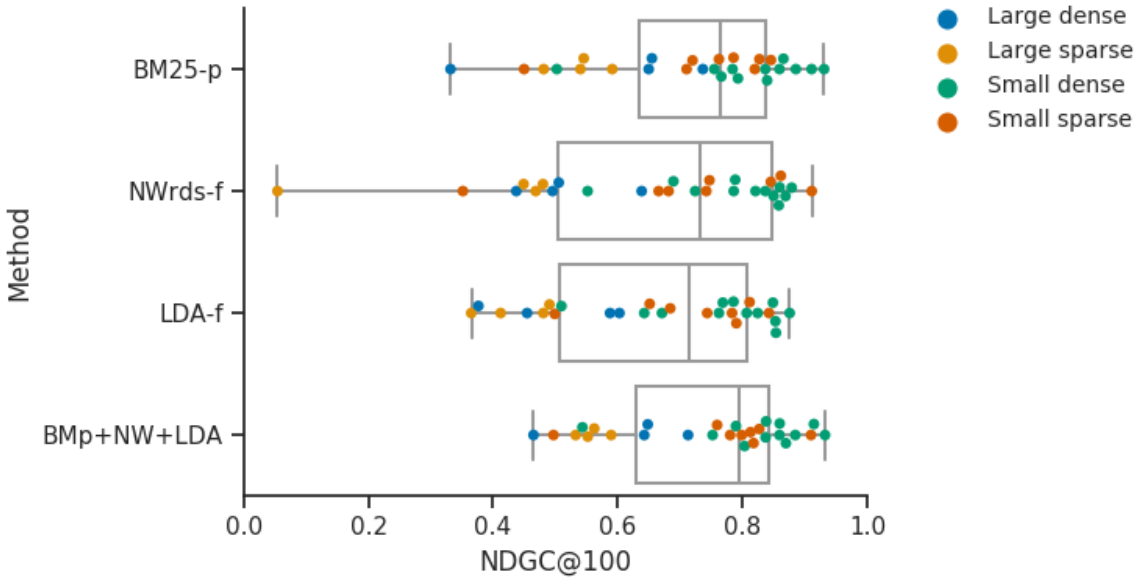

Figure 32: The figure shows scatter plots of the performance on the individual 28 queries measured in terms of NDGC@100. BMp+NW+LDA appear to perform slightly better than its constituent parts.

I adjust the p-value for 3 comparisons I reject the null with respect to LDA-f (p-value=.001). Since I rejected the null hypothesis I can step down to test NWrdsW-f, the method with the second lowest p-value (p-value=.001). After applying the adjustment ($\alpha/(k-2)$) I reject the null with respect to NWrds-f (p-value=.002). Finally, I test the BM25-p (p-value=.107) for which I cannot reject the null. Hence, I have established that the BMp+NW+LDA method performs significantly better than two of its constituent parts (NWrds-f, LDA-f).

Second, I analyze the thesis statement S6.2 by testing the corresponding null hypothesis. Again, as the first step I use the Friedman test to assess the hypothesis that at least one of the GloVe-r, NWrds-f, LDA-f, and GVr+NW+LDA methods differs from the other three. Interestingly, I cannot reject the null (p-value=.14).

The interesting properties of the methods based on novelty detection and topic similarity assessment motivated me to propose the framework that deploy these models on top of the context-aware base ranker. The goal was to apply these methods in such a way as to help the base ranker deal with the two systematic problems that we identified in [108] and that are discussed in Sections 6.3.2 and 6.4.2 (i.e., filter out cases from different domains and identify complete or partial citations of the source provision). The challenge identified in [108] was not to interfere with the base ranker too much since its ranking is vastly superior to that of the supporting methods. In [108] initial attempts to connect the models through linear interpolation yielded poor results. The approach that turned out to be successful (which I confirm here) was to apply the supporting methods as indicator variables. These only fire in case of very strong evidence that one of the problems is present and down-rank the problematic sentence.

6.7 CONCLUSIONS

I performed a detailed study on a number of ranking methods in the context of the specialized task of retrieving case-law sentences for statutory interpretation. I confirmed that retrieving the sentences directly by measuring similarity between the query and a sentence yields mediocre results. Taking into account sentences' context turned out to improve the

performance of the ranking to a certain degree. I observed that novelty detection and topic similarity assessment techniques are able to capture information that could be used as an additional layer in a ranker's decision. Based on the detailed error analysis I integrated the context-aware ranking methods with the components based on novelty detection and topic similarity assessment into a specialized framework for retrieval of case-law sentences for statutory interpretation. Evaluation of different implementations of the framework shows promising results.

7.0 FEATURE ENGINEERING-BASED LEARNING TO RANK APPROACHES FOR RETRIEVAL OF USEFUL SENTENCES

Here, I present a series of experiments that explore the use of classical feature engineering-based learning-to-rank approaches (LTR) to the task of ranking the sentences with respect to their utility. By classical LTR I mean the setup where the relationship between the query and a document is encoded in a single feature vector which is typically domain specific (e.g. LETOR 4.0 for web search [91]). The task is then to utilize this representation in a learning algorithm so that the algorithm automatically learns the ranking function. The detailed study presented in Chapter 6 enables to assemble a list of such features specifically tailored for this work. I describe the assembly of the feature list in Section 7.1. In Section 7.2 I assess the effectiveness of the straightforward approach where the task is tackled as a standard multi-label classification problem. Finally, I evaluate the performance of the same classification algorithms when the problem is transformed into a pair-wise comparison among the retrieved documents (Section 7.3).

7.1 THE FEATURES

In classical LTR one usually starts by identifying a list of features indicative of the relevance of the retrieved documents (e.g., [92, 91]). In [91] an example of a hand-crafted list of features encoding the relevance-oriented relationship between the query and a document in the context of web search is presented. These include generally applicable features such as TF-IDF scores of the various parts of the document, lengths of those parts or their language model matches to the query. Features such as inlink and outlink counts or PageRank scores

are rather specific for the web search domain. Notice that the features may pertain to various items or their interactions with other items. For example, the features such as document body or title length pertain to a retrieved document itself, while the TF-IDF score of the document or its language model match stem from the interaction between the query and the retrieved document. Furthermore, features such as the document's PageRank score or its inlink count are derived from the document's context (i.e., the links leading to the document from other documents). Here I craft such a list of 129 features based on the analysis from Chapter 6 tailored for the task of retrieving sentences that are useful for argumentation about the meaning of statutory terms. The remaining content of this section is organized by the type of item the respective features pertain too. In this section I only provide a high-level description of the individual classes of features. The full list of features with their definitions is available in Appendix D.

7.1.1 Features Based on Individual Units

This class comprises 59 features that pertain to the items themselves (e.g., the retrieved sentences, the query, the containing paragraphs). The features include low-level characteristics of the units, such as the word count, syntactic attributes, such as the POS type of the unit's root, as well as more sophisticated traits such as the item's membership in the specific functional part (Section 6.5). There are also some domain specific features such as the proportion of the unit that is covered by an explicit quotation (i.e., enclosed in quotation marks) or the count of explanatory/elaboration-oriented words. In [84, 83] the authors demonstrated success using argument keywords as indicators of argument components. Here, the principle is similar. While [84, 83] uses LDA to mine the argument words I have simply compiled the list of explanatory/elaboration-oriented words from several publicly available sources.[1] The features are listed in Section D.1.

- *Retrieved Sentence* – A retrieved sentence is represented by 11 features. (Subsection

[1] www.english-at-home.com/grammar/linking-words/; plainlanguage.gov/guidelines/words/use-simple-words-phrases/; www.macmillandictionary.com/us/thesaurus-category/american/ways-of-explaining-or-clarifying; wordvice.com/recommended-verbs-for-research-writing/; www.smart-words.org/linking-words/transition-words.html

D.1.1)

- *Query* – A query is represented by four features encoding its length and a root's POS type. (Subsection D.1.2)
- *Surrounding Sentence* – There are 28 features modeling the four sentences surrounding the retrieved sentence, i.e. the two preceding and the two following sentences. The seven features modeling each surrounding sentence are a subset of the features used to represent the retrieved sentence. They are focused on the length of the sentence, explanatory/elaboration-oriented words, and quotations. (Subsection D.1.3)
- *Paragraph* – A paragraph containing the retrieved sentence is represented by the same 7 features as each of the surrounding sentences. (Subsection D.1.4)
- *Opinion* – An opinion that contains the retrieved sentence. It is modeled by just one feature–its word count. (Subsection D.1.5)
- *Case* – A case containing the retrieved sentence is represented by its word count (i.e., one feature). (Subsection D.1.6)
- *Source Provision* – A source provision that contains the retrieved sentence is modeled by the same seven features as a paragraph or each of the surrounding sentences. (Subsection D.1.7)

7.1.2 Features Based on Matching the Query

There are 46 features that are focused on the interaction between the items and the query. The features comprise of different matching scores (e.g., BM25), the relationship of the query and quotes contained in the item, as well as various syntactic attributes related to the position of the query within the unit (e.g., subtree size). The features are listed in Section D.2.

- *Query Matched to Retrieved Sentence* – The match is represented by eight features. (Subsection D.2.1)
- *Query Matched to Surrounding Sentence* – There are 32 features modeling the match between the query and four sentences surrounding the retrieved sentence. These are the

same ones as those used for the match between the query and the retrieved sentence. (Subsection D.2.2)

- *Query Matched to Paragraph* – The match between the query and the paragraph containing the retrieved sentence is represented by two features focused on measuring its match to the query. The two features are a subset of the features used to represent the retrieved sentence. (Subsection D.2.3)
- *Query Matched to Opinion* – The match between the query and the opinion that contains the retrieved sentence. It is modeled by the same two features as the paragraph. (Subsection D.2.4)
- *Query Matched to Case* – The match between the query and the case containing the retrieved sentence is represented by the same two features as the paragraph and the opinion. (Subsection D.2.5)

7.1.3 Features Based on Matching the Source Provision

This class comprises 19 features that model the relationship between the source provision and other units. The features are focused on the topical match between the source provision and the item or the novelty of the individual unit with respect to the source provision. The features are listed in Section D.3.

- *Source Provision Matched to Retrieved Sentence* – The match is represented by two features focused on its novelty. (Subsection D.3.1)
- *Source Provision Matched to Surrounding Sentence* – There are 8 features modeling the match between the source provision and the four sentences surrounding the retrieved sentence. The two features modeling each surrounding sentence are the same ones as those used for the retrieved sentence itself. (Subsection D.3.2)
- *Source Provision Matched to Paragraph* – The match between the source provision and the paragraph containing the retrieved sentence is represented by four features focused on the topical match as well as its novelty. (Subsection D.3.3)
- *Source Provision Matched to Opinion* – The match between the source provision and the opinion that contains the retrieved sentence. It is modeled by two features focused on

its topical match to the source provision. (Subsection D.3.4)

- *Source Provision Matched to Case* – The match between the source provision and the case containing the retrieved sentence is represented by three features focused on its topical match. (Subsection D.3.5)

7.1.4 Features Based on the List of Results

There are five features that model certain aspects of the retrieved results list. These features correspond to the statistics reported in Table 6 (i.e., number of retrieved sentences, sentence/case ratio, etc.). They are listed in Section D.4.

7.2 POINT-WISE APPROACH

The simplest LTR approach to ranking is to represent each document as a single feature vector. Given the existence of ground-truth labels (four ordinal categories in our case) one can treat the problem as a standard classification or regression task. The goal is to learn a function that takes the feature vector of a document as input and predicts its relevance degree. Based on such function one can sort the documents into a ranked list. The ranking may be modeled as regression, classification, or ordinal regression/classification. [66] Almost any standard regression or classification algorithm can be used and many have been tried:[2]

- Linear regression [17]
- SVM [115, 116]
- Logistic regression [40]
- Decision tree ensembles [65]
- Neural networks [18]

In this section I report the results of the experiments on applying several of the regression and classification methods. These receive the features described in Section 7.1 as the input and based on those they learn to predict sentences' value (4 level ordinal scale).

[2]The list of algorithms and references is reported in [66].

7.2.1 Experiments

The experimental setup is similar to the one used in Chapter 6. I first retrieve all the sentences that contain the term of interest. The individual sentences are then represented as features described in Section 7.1. Using different off-the-shelf ML algorithms, sentences are ranked from the ones predicted as the most valuable to those deemed the least valuable. As baselines I report the performance of the BM25 (Section 6.1) and BM25-c (Section 6.2) methods as well as the performance of the Random system (defined in Section 6.1.1) for reference.

The one difference in the experimental setup, compared to the experiments from Chapter 6, is the inclusion of the second and sixth folds. The two folds are added to be used as a held-out set. These were not used in Chapter 6. Hence, the experiments in this chapter are performed on all the 42 queries as opposed to just 28 in Chapter 6. Each method is evaluated in terms of two experimental setups. The first setup is a cross-validation where in each step the algorithm uses one of the six folds as a test set. The remaining folds are used for training and optimization of hyperparameters (validation). Each fold serves as a test set exactly once. Therefore, I perform a 6-fold nested cross-validation where in each run a 5-fold cross-validation is performed on the training set (5 folds) to optimize the hyperparameters. The second setup focuses on performance of the method on the second and sixth folds that were originally held out. The principle is otherwise the same, i.e., in each of the two steps the other 5 folds serve as the training and validation set. The rationale behind using the first setup is to obtain as robust a set of results as possible (refer back to Section 5 where it is explained that in IR, train-test split matters). The reason for using the second setup is to make sure that the approach generalizes to queries that were not considered during the experiments in Chapter 6. This is important because in Chapter 6, I performed the analysis on the basis of which the feature list (Section 7.1) was put together. By using the held-out set I make sure the features generalize to unseen queries.

The first class of models I employ are regression-based algorithms. In this setup the relevance of a document (4 degree ordinal scale in our case) is regarded as a continuous variable. [66] Despite the discrepancy between the nature of labels a regression algorithm

expects (continuous) and the nature of labels I provide (ordinal), the approach is appealing. In [17] it is shown that the square loss could be understood as an upper bound to the NDGC-based ranking error. Even if there is a considerable regression loss, the ranking can still be reasonable as long as the relative orderings between the predictions are mostly correct. [66] I evaluate the linear regression model with L2 regularization.[3] The prediction $\hat{y}_i$ is understood as the respective sentence's score, which determines the order of the sentence in the ranking. One more regression model I experiment with is an AdaBoost regressor [35, 25] with decision tree regressor as the base estimator [13]. AdaBoost first fits the base regressor to the whole dataset and then creates its copies that focus more on difficult cases.[4]

The second class of models I analyze are classification-based algorithms. In this setup the relevance of a document is regarded as a categorical variable. Here the discrepancy between the nature of labels is different. The algorithm no longer treats them as continuous, which is preferable, but the information about their ordering is lost, which is a disadvantage. From the perspective of the standard classification algorithm, a prediction of the 'no value' label instead of a 'high value' is the same kind of mistake as predicting the 'certain value' label. In prediction I obtain the per-class probability vector ($\vec{p_i}$) for sentence s_i:

$$(p(s_i = \text{'no value'}), p(s_i = \text{'potential value'}), p(s_i = \text{'certain value'}), p(s_i = \text{'high value'}))$$

To obtain the sentence's score I compute an inner product between $\vec{p}$ and value weight vector $\vec{w}$:

$$\vec{p}(0, 1, 2, 3)^T$$

The motivation to use this approach over considering the predicted class only is to take into account the confidence of the prediction. Note that this may lead to sentences that are being predicted as having higher value to be ranked below some sentences that were predicted as having lower value as shown in the following example:

$$(0.0, 0.0, 0.51, 0.49)\vec{w} = 2.49 > 1.51 = (0.25, 0.25, 0.24, 0.26)\vec{w}$$

[3] scikit-learn.org/stable/modules/generated/sklearn.linear_model.Ridge.html
[4] scikit-learn.org/stable/modules/generated/sklearn.ensemble.AdaBoostRegressor.html

Here, the first sentence would be ranked higher, despite being predicted as having only 'certain value,' than the second sentence which is predicted as having 'high value.'

In the experiments I use a logistic regression model with L2 regularization.[5] [40] One more linear model I evaluate is support vector machines (SVM). [115, 116] An SVM classifier constructs a hyper-plane in a high dimensional space, which is used to separate the classes from each other. As an implementation of SVM, I used the scikit-learn's Support Vector Classification module.[6] From the non-linear models I work with a random forest. A random forest is an ensemble classifier (as is the one used in [65]) that fits a number of decision trees on sub-samples of the data set. It uses averaging to improve the predictive accuracy and control over-fitting. As an implementation of random forest, I used the scikit-learn's Random Forest Classifier module.[7] Finally, I evaluated the performance of a multi-layer perceptron classifier.[8] It differs from the logistic regression by including one or more hidden layers between the input and output layers.

Finally, I use the same models in a setup where the ordinal nature of the labels is being taken into account. To achieve this goal I use the method from [34]. In this setup the task is transformed from a single k-class classification problem to $k-1$ binary classification problems. In this work it means I create three sets of labels using the following transformation functions:

$$f_1(s_i) = \begin{cases} 0 & \text{if } s_i \text{ has no value} \\ 1 & \text{else} \end{cases}$$

$$f_2(s_i) = \begin{cases} 0 & \text{if } s_i \text{ is in } \{\text{no value, potential value}\} \\ 1 & \text{else} \end{cases}$$

$$f_3(s_i) = \begin{cases} 1 & \text{if } s_i \text{ has high value} \\ 0 & \text{else} \end{cases}$$

[5] scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

[6] scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html

[7] scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

[8] scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html

Intuitively, f_1 encodes which sentences have a higher value label than 'no value.' Similarly, f_2 encodes which of them have a higher value than 'potential value' and f_3 does the same for 'certain value.' Sequential application of the 3 classifiers (each of them trained on one of the transformations) yields the probability distribution over the 4 classes:

$$\begin{aligned}
p(s_i = \text{'no value'}) &= 1 - p(s_i > \text{'no value'}) \\
p(s_i = \text{'potential value'}) &= p(s_i > \text{'no value'}) - p(s_i > \text{'potential value'}) \\
p(s_i = \text{'certain value'}) &= p(s_i > \text{'potential value'}) - p(s_i > \text{'certain value'}) \\
p(s_i = \text{'high value'}) &= p(s_i > \text{'certain value'})
\end{aligned}$$

Using this approach the ordinal nature of the labels is respected and therefore, I would expect an increase in classification performance (as was shown in [34]). As the classification performance somewhat correlates with the ranking performance, the hope is that using this method will improve the quality of the rankings.

7.2.2 Results and Discussion

The results of the experiments with regression and multi-label classification models described in Section 7.2.1 are reported in Table 18 (group and overall means) and Figure 33 (box and whisker plots and swarm plots). The two regression and the four classification methods trained on the features described in Section 7.1 all appear to perform better than the two baselines.

First, I examine the thesis statement S7.1 by testing the corresponding null hypothesis. As the first step I use the Friedman test to assess the hypothesis that at least one of the BM25, BM25-c, LinReg, AdaBoost, LogReg, SVM, RF, and MLP methods differs from the other 7. The methods are compared in terms of their performance as measured with NDGC@100 on the Overall. I reject the hypothesis for both the cross-validation (p-value=.0002) as well as for the held-out set setup (p-value=.035). Using the Holm procedure I recognize the

Method	SmSp @10	SmSp @100	SmDs @10	SmDs @100	LgSp @10	LgSp @100	LgDs @10	LgDs @100	Overall @10	Overall @100
CROSS-VALIDATION										
Random	.38 ± .10	.67 ± .15	.52 ± .07	.76 ± .09	.25 ± .16	.29 ± .18	.47 ± .09	.48 ± .09	.43 ± .13	.63 ± .21
BM25	.47 ± .13	.74 ± .11	.60 ± .18	.79 ± .11	.44 ± .21	.37 ± .22	.61 ± .17	.56 ± .12	.54 ± .18	.68 ± .20
BM25-c	.48 ± .12	.76 ± .09	.59 ± .17	.80 ± .11	.49 ± .14	.42 ± .17	.63 ± .19	.55 ± .13	.55 ± .16	.70 ± .18
LinReg	.55 ± .17	.77 ± .12	.65 ± .15	.82 ± .10	.67 ± .08	.60 ± .08	.55 ± .12	.61 ± .08	.61 ± .15	.75 ± .14
AdaBoost	.57 ± .15	.79 ± .09	.62 ± .12	.81 ± .11	.63 ± .11	.61 ± .08	.63 ± .15	.64 ± .08	.61 ± .13	.75 ± .13
LogReg	.56 ± .13	.78 ± .10	.67 ± .12	.83 ± .10	.62 ± .11	.57 ± .10	.70 ± .20	.65 ± .13	.64 ± .14	.75 ± .14
SVM	.59 ± .16	.80 ± .11	.66 ± .14	.84 ± .11	.64 ± .16	.63 ± .05	.63 ± .13	.63 ± .13	.63 ± .14	**.77 ± .13**
RF	.61 ± .18	.81 ± .11	.66 ± .16	.83 ± .11	.56 ± .19	.55 ± .17	.54 ± .20	.58 ± .15	.61 ± .18	**.75 ± .17**
MLP	.58 ± .16	.78 ± .10	.66 ± .13	.82 ± .09	.65 ± .08	.59 ± .10	.60 ± .21	.63 ± .15	.63 ± .15	**.75 ± .14**
HELD-OUT SET										
Random	.34 ± .14	.65 ± .16	.52 ± .05	.75 ± .07	.16 ± .18	.17 ± .17	.47 ± .03	.49 ± .00	.41 ± .16	.60 ± .23
BM25	.41 ± .04	.73 ± .09	.57 ± .20	.77 ± .10	.40 ± .45	.14 ± .07	.56 ± .18	.53 ± .16	.50 ± .20	.63 ± .24
BM25-c	.50 ± .06	.76 ± .05	.55 ± .17	.79 ± .08	.46 ± .28	.23 ± .01	.56 ± .16	.52 ± .13	.52 ± .15	.66 ± .22
LinReg	.57 ± .18	.80 ± .10	.64 ± .15	.82 ± .10	.72 ± .11	.65 ± .07	.50 ± .06	.58 ± .06	.61 ± .15	.76 ± .13
AdaBoost	.56 ± .14	.81 ± .08	.64 ± .12	.79 ± .13	.72 ± .11	.59 ± .16	.49 ± .03	.61 ± .01	.61 ± .13	.74 ± .13
LogReg	.57 ± .17	.79 ± .10	.69 ± .08	.82 ± .09	.63 ± .24	.59 ± .06	.58 ± .09	.58 ± .06	.63 ± .13	.75 ± .13
SVM	.63 ± .19	.83 ± .12	.70 ± .13	.84 ± .10	.66 ± .25	.69 ± .02	.62 ± .01	.58 ± .03	.66 ± .14	**.78 ± .13**
RF	.60 ± .15	.84 ± .08	.69 ± .17	.83 ± .12	.41 ± .27	.39 ± .24	.42 ± .08	.51 ± .09	.58 ± .19	**.73 ± .21**
MLP	.57 ± .16	.77 ± .13	.69 ± .09	.82 ± .07	.61 ± .11	.60 ± .02	.45 ± .29	.53 ± .13	.61 ± .15	.73 ± .14

Table 18: The table shows the results of the experiments with point-wise LTR methods. The NDGC@10 and NDGC@100 are shown for the small sparse queries (SmSp), small dense queries (SmDs), large sparse queries (LgSp), large dense queries (LgDs), and all of them together (Overall).

following LTR methods as significantly different from the two baselines after the adjustment in the cross-validation experiment:

- SVM from BM25 (p-value=.0008) and from BM25-c (p-value=.01)
- RF from BM25 (p-value=.0006) and from BM25-c (p-value=.009)
- MLP from BM25 (p-value=.0006) and from BM25-c (p-value=.008)
- LinReg from BM25 (p-value=.007)
- LogReg from BM25 (p-value=.002)
- AdaBoost from BM25 (p-value=.014)

Using the Holm procedure I recognize the following LTR methods as significantly different from the two baselines after the adjustment in the held-out set settings:

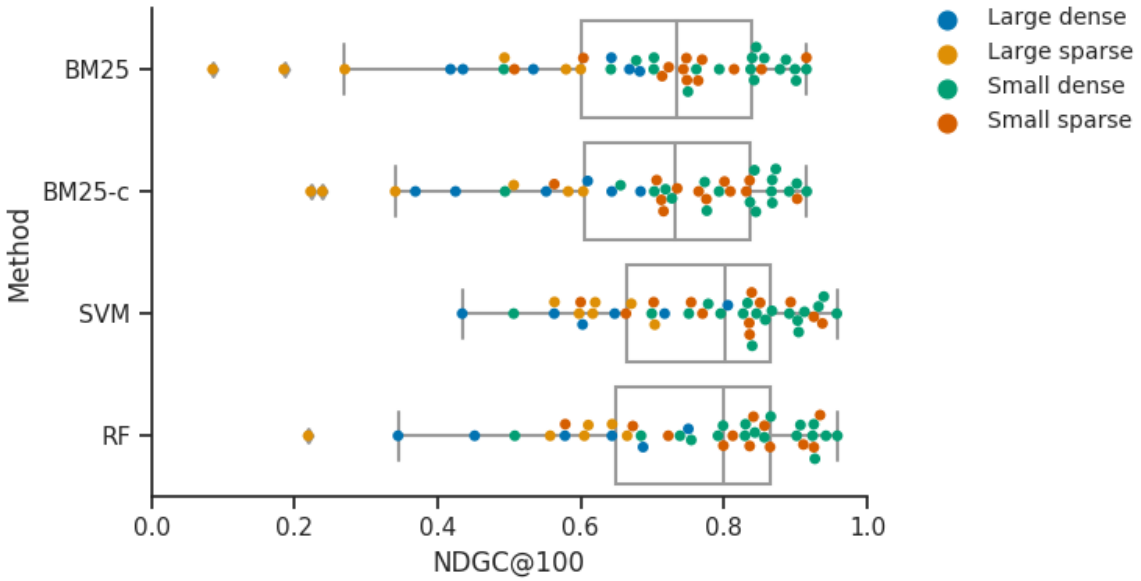

Figure 33: The figure shows scatter plots of the performance on the individual 42 queries measured in terms of NDGC@100. SVM-based and RF-based rankers perform better than the BM25 and BM25-c baselines.

- SVM from BM25 (p-value=.0008) and from BM25-c (p-value=.038)
- RF from BM25 (p-value=.0047) and from BM25-c (p-value=.033)

Hence, I conclude that SVM and RF-based rankers showed significantly different performance under both experimental settings. For the MLP based ranker I was able to establish the difference only in the cross-validation setup.

The results of the experiments with ordinal classification models described in Section 7.2.1 are reported in Table 19. Here, the four ordinal classification methods are compared to the SVM multi-label classification model which is the most successful one from the previous batch of experiments. Overall, the performance does not seem to be improved over the base method.

I analyze the thesis statement S7.2 by testing the corresponding null hypothesis. Again, as the first step I use the Friedman test to assess the hypothesis that at least one of the SVM, LogReg-ORD, SVM-ORD, RF-ORD, and MLP-ORD methods differs from the other four.

Method	SmSp @10	SmSp @100	SmDs @10	SmDs @100	LgSp @10	LgSp @100	LgDs @10	LgDs @100	Overall @10	Overall @100
CROSS-VALIDATION										
SVM	.59 ± .16	.80 ± .11	.66 ± .14	.84 ± .11	.64 ± .16	.63 ± .05	.63 ± .13	.63 ± .13	.63 ± .14	.77 ± .13
LogReg-ORD	.55 ± .17	.77 ± .11	.64 ± .15	.82 ± .10	.73 ± .09	.64 ± .04	.66 ± .16	.64 ± .11	.63 ± .15	.76 ± .12
SVM-ORD	.56 ± .15	.80 ± .09	.65 ± .12	.82 ± .09	.62 ± .13	.61 ± .04	.72 ± .20	.64 ± .13	.63 ± .15	.76 ± .13
RF-ORD	.63 ± .20	.82 ± .10	.64 ± .16	.83 ± .10	.71 ± .18	.67 ± .08	.59 ± .15	.58 ± .14	.64 ± .17	.77 ± .14
MLP-ORD	.53 ± .23	.76 ± .14	.64 ± .13	.82 ± .10	.61 ± .10	.59 ± .06	.65 ± .18	.61 ± .15	.60 ± .17	.74 ± .15
HELD-OUT SET										
SVM	.63 ± .19	.83 ± .12	.70 ± .13	.84 ± .10	.66 ± .25	.69 ± .02	.62 ± .01	.58 ± .03	.66 ± .14	.78 ± .13
LogReg-ORD	.58 ± .17	.80 ± .11	.65 ± .15	.81 ± .11	.77 ± .05	.68 ± .01	.58 ± .03	.60 ± .04	.64 ± .14	.76 ± .12
SVM-ORD	.56 ± .18	.81 ± .11	.67 ± .11	.82 ± .09	.61 ± .10	.57 ± .04	.63 ± .06	.58 ± .07	.63 ± .12	.75 ± .14
RF-ORD	.63 ± .19	.84 ± .09	.67 ± .13	.83 ± .09	.87 ± .05	.75 ± .10	.46 ± .12	.53 ± .12	.66 ± .17	.78 ± .14
MLP-ORD	.60 ± .15	.80 ± .07	.68 ± .10	.83 ± .09	.61 ± .19	.62 ± .06	.60 ± .17	.57 ± .07	.64 ± .12	.76 ± .13

Table 19: The table shows the results of the experiments with point-wise LTR methods with ordinal classifiers. The NDGC@10 and NDGC@100 are shown for the small sparse queries (SmSp), small dense queries (SmDs), large sparse queries (LgSp), large dense queries (LgDs), and all of them together (Overall).

The methods are compared in terms of their performance as measured with NDGC@100 on the Overall. I fail to reject the hypothesis for both, the cross-validation setup (p-value=.204) as well as for the held-out set (p-value=.267). In this case, the outcome seems to correspond to what we observe in Table 19. Therefore, I conclude that casting the task as an ordinal classification problem does not appear to result in improved ranking performance. Interestingly, I have observed quite a noticeable improvement on the classification performance. However, it appears that although these two are related, they do not correlate closely.

7.3 PAIR-WISE APPROACH

The point-wise approach to ranking (Section 7.2) has some intrinsic problems that are very well-known. As pointed out in [66] the first major shortcoming is the blindness of the model to the fact that the retrieved documents are grouped by the respective queries. Consequently, as the number of documents associated with different queries varies widely some of the

queries have much more weight than others (e.g., 'mechanical recordation' is associated with 18 sentences whereas 'useful improvement' has 3,868). Another problem is the fact that the training optimizes for the regression/classification performance, whereas what matters is the relative ordering of the documents. [66] Basically, even though two documents are misclassified, the outcome is optimal as long as the less valuable document is scored below the more valuable one. On the other hand, as I confirmed in Section 7.2, higher classification performance does not automatically translate into higher quality ranking.

The second of the above described problems is mitigated by the pair-wise approach, which focuses on the relative order between two documents. Here, the algorithms are trained to assess a pair of documents and decide which of the two should be placed higher. Hence, the pair-wise approach is closer to the concept of "ranking" than the point-wise approach. [66] The core idea behind the pair-wise approach is the following transformation of the data set:

$$\langle \mathbf{X}, \mathbf{y} \rangle = \left\langle \begin{array}{ccc|c} x_{1,1} & \cdots & x_{1,m} & y_1 \\ \vdots & \ddots & \vdots & \vdots \\ x_{n,1} & \cdots & x_{n,m} & y_n \end{array} \right\rangle \rightarrow \left\langle \begin{array}{ccc|c} x_{1,1} - x_{i,1} & \cdots & x_{1,m} - x_{i,m} & y_1 > y_i \\ \vdots & \ddots & \vdots & \vdots \\ x_{n,1} - x_{j,1} & \cdots & x_{n,m} - x_{j,m} & y_n > y_j \end{array} \right\rangle$$

Here, $x_{i,j}$ stands for the j-th feature of the i-th document $\vec{x_i} \in \mathbf{X}$. The two notable outcomes of the transformation are the following:

1. The number of data points in the data set increases from n to at most $n(n-1)/2$.
2. The original labels (continuous, ordinal, binary) become binary (i.e., 'more valuable,' 'less valuable').

The increase in the size of the data set could be beneficial for small data sets but it could also be a concern when it comes to larger data sets. For example, in this work plugging the total number of sentences (i.e., 26,959) into the formula I would get 363,380,361 data points. However, in the context of ad hoc information retrieval it is customary to only consider the pairs of documents which are associated with the same query. The data set in this work has smaller document sets associated with the queries. The largest one is the list corresponding to the 'useful improvement' query. Since the list has 3,867 sentences the upper bound for the number of data points resulting from this transformation is 7,474,911. Since the pairs of sentences with the same value are discarded, the actual number would typically be less

than half of the upper bound. Hence, the potential scalability issue does not manifest itself in the data set in this work. This is why I do not account for it. However, it would be important to keep this issue in mind and implement a coping strategy in case it would be needed. A typical approach is to retrieve a limited set of documents with a less powerful but more scalable method and apply the transformation to that limited set only.

The transformation of the original problem into a binary classification one does not pose any troubles. The big upside of this transformation is that, since each pair of sentences is considered only once, one can choose which of the sentences will be the reference one (i.e., the first in the pair). By doing this one can control which label will result (i.e., 'less valuable' or 'more valuable'). As a consequence one can always obtain a perfectly balanced data set no matter the distribution of the original labels.

7.3.1 Experiments

The experimental setup is the same as the one used in Section 7.2.1. The only difference is that the transformation of the data set described above is used, i.e., the task is transformed into the pair-wise binary classification. The sentence pairs are then represented in terms of the difference of the features described in Section 7.1. Using the same off-the-shelf classification algorithms as before (logistic regression, support vector machines, random forest, multi-layer perceptron) sentences are ranked from the ones predicted as the most valuable to those deemed the least valuable. As baselines I report the performance of the RF-ORD, the most powerful system from Section 7.2.2 where the point-wise approach was applied. As in the experiments in Section 7.2.1 all six folds, i.e., all the 42 queries, are included. Again, each method is evaluated in terms of the cross-validation and the held-out set experimental setups (for more details refer to Section 7.2.2).

In prediction for each sentence pair that is considered I obtain the probability of the first sentence being more valuable. The contribution of each sentence pair (s_i, s_j) to the overall aggregate score is shown below:

$$s_i \leftarrow p(s_i > s_j) - p(s_i < s_j)$$

$$s_j \leftarrow p(s_i < s_j) - p(s_i > s_j)$$

Method	SmSp @10	SmSp @100	SmDs @10	SmDs @100	LgSp @10	LgSp @100	LgDs @10	LgDs @100	Overall @10	Overall @100
CROSS-VALIDATION										
RF-ORD	.63 ± .20	.82 ± .10	.64 ± .16	.83 ± .10	.71 ± .18	.67 ± .08	.59 ± .15	.58 ± .14	.64 ± .17	.77 ± .14
LR-PWT	.58 ± .21	.79 ± .13	.64 ± .16	.82 ± .11	.59 ± .09	.56 ± .10	.72 ± .15	.65 ± .08	.63 ± .17	.75 ± .14
SVM-PWT	.56 ± .15	.80 ± .09	.65 ± .12	.82 ± .09	.62 ± .13	.61 ± .04	.72 ± .20	.64 ± .13	.63 ± .15	.76 ± .13
RF-PWT	.60 ± .16	.81 ± .11	.66 ± .12	.83 ± .10	.71 ± .17	.68 ± .08	.67 ± .10	.64 ± .09	.65 ± .14	.77 ± .12
MLP-PWT	.53 ± .23	.76 ± .14	.64 ± .13	.82 ± .10	.61 ± .10	.59 ± .06	.65 ± .18	.61 ± .15	.60 ± .17	.74 ± .15
HELD-OUT SET										
RF-ORD	.63 ± .19	.84 ± .09	.67 ± .13	.83 ± .09	.87 ± .05	.75 ± .10	.46 ± .12	.53 ± .12	.66 ± .17	.78 ± .14
LR-PWT	.54 ± .19	.81 ± .08	.63 ± .17	.81 ± .12	.69 ± .00	.63 ± .11	.59 ± .13	.61 ± .09	.61 ± .15	.75 ± .13
SVM-PWT	.56 ± .18	.81 ± .11	.67 ± .11	.82 ± .09	.61 ± .10	.57 ± .04	.63 ± .06	.58 ± .07	.63 ± .12	.75 ± .14
RF-PWT	.59 ± .18	.83 ± .08	.69 ± .11	.83 ± .08	.87 ± .02	.77 ± .09	.61 ± .02	.60 ± .04	.67 ± .14	.79 ± .11
MLP-PWT	.60 ± .15	.80 ± .07	.68 ± .10	.83 ± .09	.61 ± .19	.62 ± .06	.60 ± .17	.57 ± .07	.64 ± .12	.76 ± .13

Table 20: The table shows the results of the experiments with pair-wise LTR methods. The NDGC@10 and NDGC@100 are shown for the small sparse queries (SmSp), small dense queries (SmDs), large sparse queries (LgSp), large dense queries (LgDs), and all of them together (Overall).

The motivation to use this approach over simply considering the number of "wins," as is usually done, is to take into account the confidence of the prediction. Consider the following two examples:

$$p(s_i > s_j) = 1.0$$

$$p(s_i > s_j) = .51$$

In both cases, s_i is deemed more valuable. Whereas in the first case the contribution of this comparison is going to be $s_i \leftarrow 1$ and $s_j \leftarrow -1$, the contribution in the second case is only $s_i \leftarrow .02$ and $s_j \leftarrow -.02$. Using this strategy I minimize the influence of data points (sentence pairs) where the classifier has low confidence in its prediction.

7.3.2 Results and Discussion

The results of the experiments described in Section 7.3.1 are reported in Table 20 (group and overall means). Here, the four pair-wise classification methods are compared to the

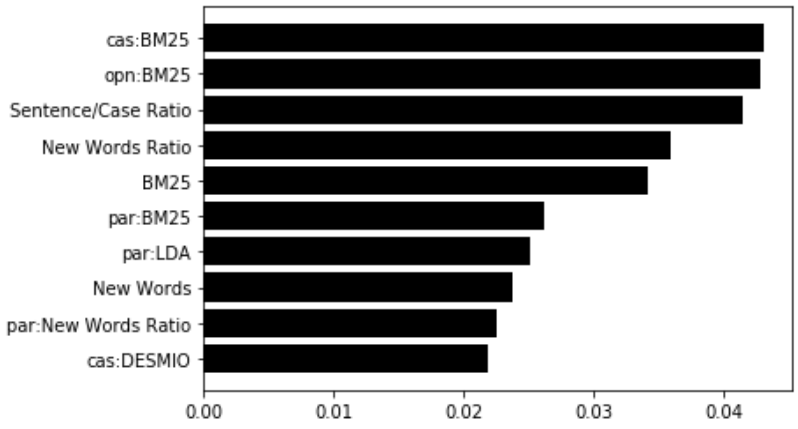

Figure 34: The figure shows the most important features as recognized by the Random Forest classifier.

RF-ORD multi-label classification model which is the most successful one from the previous batches of experiments. Overall, the performance does not seem to be improved over the base method.

I analyze the thesis statement S7.3 by testing the corresponding null hypothesis. Again, as the first step I use the Friedman test to assess the hypothesis that at least one of the RF-ORD, LR-PWT, SVM-PWT, RF-PWT, and MLP-PWT methods differs from the other four. The methods are compared in terms of their performance as measured with NDGC@100 on the Overall. Again, I fail to reject the hypothesis for both, the cross-validation setup (p-value=.217) as well as for the held-out set (p-value=.183). The outcome seems to correspond to what we observe in Table 20. Therefore, I conclude that casting the task as pair-wise classification problem does not appear to manifest in improved ranking performance.

7.4 ABLATION STUDY

Since I have established that learning-to-rank approaches result in methods producing sophisticated rankings outperforming the BM25 and the BM25-c baselines, I would be interested

to understand what features are the most useful ones. For this analysis I will use the simple Random Forest model with 100 base estimators. Figure 34 shows the 10 most important features when the algorithm is trained on all of the 129 features. Interestingly, the two baseline methods (BM25 and BM25-c) are among the most useful features. Furthermore, all of the features (except one) are those that encode the relationship between either the term of interest or the source provision and a sentence. The one exception is the Sentence/Case Ratio feature which appears to inform the classifier about an important property of a query. Note that this feature could be understood as a proxy for the richness of a query as described in Section 5.1. This seems to suggest that different queries may require different treatment—an interesting insight for future work. The features that describe the sentence or its context itself do not seem to be that useful.

The "single group" portion of Table 21 shows the results of the experiments when only a small group of features modeling certain phenomenon is considered. I have grouped the features into the following classes:

- *sentence* – The features that describe the retrieved sentence itself.
- *surrounding* – These features model the immediate surroundings of the retrieved sentence (i.e. the 2 preceding and the 2 following sentences). Importantly this group of features does not include and is completely unaffected by the retrieved sentence itself.
- *par-opn-case* – The features encode basic characteristics of the higher textual units that contain the retrieved sentence, i.e., the containing paragraph, the opinion, and the case.
- *qry2snt* – These features describe the match between the term of interest (the query) and the retrieved sentence.
- *qry2surr* – The features model the match between the term of interest and the immediate surroundings of the retrieved sentence.
- *qry2poc* – These features encode the match between the term of interest and the higher textual units that contain the retrieved sentence.
- *sp2snt* – These features describe the match between the source provision and the retrieved sentence.
- *sp2surr* – The features model the match between the source provision and the immediate surroundings of the retrieved sentence.

Method	SmSp @10	SmSp @100	SmDs @10	SmDs @100	LgSp @10	LgSp @100	LgDs @10	LgDs @100	Overall @10	Overall @100
Random	.38 ± .10	.67 ± .15	.52 ± .07	.76 ± .09	.25 ± .16	.29 ± .18	.47 ± .09	.48 ± .09	.43 ± .13	.63 ± .21
RF	.61 ± .21	.81 ± .12	.61 ± .14	.81 ± .10	.63 ± .16	.59 ± .07	.69 ± .19	.62 ± .12	.63 ± .17	.75 ± .14
SINGLE GROUP										
sentence	.43 ± .16	.71 ± .14	.52 ± .16	.76 ± .12	.30 ± .12	.32 ± .15	.49 ± .11	.48 ± .11	.46 ± .16	.64 ± .20
surrounding	.39 ± .12	.67 ± .15	.49 ± .13	.75 ± .12	.22 ± .14	.27 ± .16	.43 ± .11	.47 ± .11	.41 ± .15	.62 ± .22
par-opn-cas	.39 ± .16	.67 ± .17	.49 ± .13	.75 ± .10	.24 ± .16	.28 ± .19	.43 ± .14	.45 ± .10	.41 ± .16	.61 ± .22
qry2snt	.46 ± .10	.73 ± .13	.56 ± .10	.78 ± .07	.30 ± .24	.31 ± .23	.38 ± .20	.50 ± .06	.47 ± .17	.66 ± .21
qry2surr	.47 ± .13	.73 ± .14	.49 ± .11	.76 ± .10	.23 ± .18	.32 ± .16	.53 ± .14	.50 ± .09	.45 ± .16	.65 ± .20
qry2poc	.32 ± .11	.64 ± .17	.58 ± .12	.79 ± .09	.26 ± .17	.33 ± .19	.44 ± .29	.48 ± .14	.44 ± .20	.64 ± .21
sp2snt	.57 ± .18	.80 ± .12	.53 ± .16	.77 ± .11	.44 ± .18	.49 ± .08	.50 ± .17	.53 ± .10	.52 ± .17	.70 ± .16
sp2surr	.42 ± .12	.71 ± .13	.50 ± .10	.75 ± .09	.27 ± .19	.34 ± .16	.40 ± .17	.48 ± .12	.43 ± .15	.64 ± .19
sp2poc	.55 ± .20	.78 ± .14	.58 ± .13	.80 ± .11	.46 ± .12	.51 ± .08	.65 ± .13	.59 ± .12	.56 ± .15	.72 ± .16
LEAVE ONE GROUP OUT										
-sentence	.55 ± .16	.80 ± .10	.64 ± .15	.83 ± .09	.64 ± .11	.63 ± .06	.61 ± .17	.61 ± .14	.61 ± .15	.76 ± .13
-surrounding	.63 ± .22	.82 ± .13	.67 ± .14	.83 ± .11	.65 ± .17	.63 ± .07	.67 ± .20	.63 ± .15	.65 ± .17	.77 ± .15
-par-opn-cas	.61 ± .19	.81 ± .10	.64 ± .16	.82 ± .11	.62 ± .17	.61 ± .05	.66 ± .22	.63 ± .13	.63 ± .17	.76 ± .13
-qry2snt	.61 ± .22	.82 ± .12	.63 ± .13	.81 ± .11	.58 ± .15	.60 ± .05	.65 ± .19	.61 ± .15	.62 ± .17	.75 ± .15
-qry2surr	.58 ± .21	.80 ± .13	.68 ± .15	.84 ± .10	.65 ± .08	.62 ± .06	.57 ± .17	.60 ± .12	.63 ± .17	.76 ± .15
-qry2poc	.59 ± .23	.80 ± .13	.64 ± .16	.82 ± .11	.65 ± .12	.64 ± .03	.65 ± .19	.63 ± .14	.63 ± .18	.76 ± .14
-sp2snt	.53 ± .14	.78 ± .12	.65 ± .15	.82 ± .09	.57 ± .16	.58 ± .07	.67 ± .15	.60 ± .11	.61 ± .15	.74 ± .14
-sp2surr	.62 ± .19	.82 ± .12	.67 ± .18	.83 ± .10	.68 ± .14	.61 ± .07	.71 ± .16	.63 ± .13	.66 ± .17	.77 ± .14
-sp2poc	.55 ± .14	.79 ± .10	.62 ± .14	.82 ± .09	.64 ± .13	.57 ± .08	.63 ± .26	.60 ± .17	.60 ± .16	.74 ± .14
-results list	.61 ± .21	.81 ± .12	.65 ± .13	.83 ± .10	.57 ± .17	.57 ± .15	.63 ± .19	.61 ± .12	.63 ± .17	.76 ± .16
sp2snt+poc	.57 ± .20	.79 ± .13	.59 ± .14	.80 ± .11	.60 ± .07	.60 ± .04	.61 ± .09	.59 ± .12	.59 ± .14	.74 ± .14
qry+sp2snt+poc	.59 ± .18	.80 ± .12	.67 ± .14	.83 ± .10	.58 ± .21	.56 ± .18	.63 ± .23	.63 ± .12	.63 ± .18	.75 ± .16

Table 21: The table shows the results of the ablation study. The upper portion (single group) shows the performance on the individual group of features. The lower portion (leave one group out) shows the performance on the remaining features if the group is removed. The NDGC@10 and NDGC@100 are shown for the small sparse queries (SmSp), small dense queries (SmDs), large sparse queries (LgSp), large dense queries (LgDs), and all of them together (Overall).

- *sp2poc* – These features encode the match between the source provision and the higher textual units that contain the retrieved sentence.
- *result list* – The features describe the list of results (e.g., its length, the sentence/case ratio). Note that this feature set is considered only in the leave-one-group-out setup since it does not make sense to use it in the single-group setup.

Note that for reference Table 21 shows the performance of the Random system as well as the Random Forest-based ranker using all the 129 features. When the systems based on the

specialized feature groups are compared to the two reference methods it appears that only a small number of groups carry a stronger signal that is useful for ranking. Specifically, it appears that only the *sp2snt* features as well as the *sp2poc* features are capable to support the learning of the system that is clearly better than the Random baseline. Perhaps, the *qry2snt* appear to carry certain signal. Otherwise, the performance of the systems seem to be fairly close to the Random baseline.

A completely different view of the usefulness of the features is offered by the bottom part of Table 21. Here, I report the performance of the systems that are trained on all the features but the ones coming from a specific group. Interestingly, leaving out a certain group does not have much of an effect on the performance of the system. This suggests that the feature set is robust and many features model similar phenomena. This observation is corroborated by the fact that the performance even appears to improve slightly when certain groups are omitted (e.g., the *surrounding* or the *sp2surr* groups). This is a clear sign that some of the groups model similar phenomena (at least partially).

The above analysis is interesting with respect to future work since it appears that feature selection would be beneficial. A first stab in this direction is provided in the lowest part of Table 21. Here, I first evaluate the system that uses the features based on matching the source provision to the retrieved sentence and the higher textual units in which it is contained. Even though these two groups appear to be the only ones that are really useful, the performance appears to be lower than that of the system which is trained on the full feature set. This is especially true when the system is evaluated in terms of the NDGC@10 (.59 vs .63). However, the system trained on these two groups and the group that encodes the relationship between the query and the sentence appears to be competitive with the system trained on all the features.

7.5 CONCLUSIONS

In this chapter I assembled a list of 129 features modeling the term of interests, the source provisions, the retrieved sentences, their contexts, as well as the relationships among these

constituents. These features were largely based on the analysis performed in Chapter 6. I showed that these features support several learning algorithms that are capable to fit a multi-class classification or regression function that serves as a basis of a sophisticated ranking function outperforming the BM25 and BM25-c baselines. I experimented with casting the task into ordinal classification (4 ordered categories) as well as into pair-wise binary document classification (more relevant vs less relevant). Despite a notable improvement in classification performance (in the ordinal setup) I did not observe a notable increase in ranking performance. The ablation study revealed that the most useful features are those focused on modeling the relationship between the source provision and the retrieved sentences as well as with the higher level textual units in which they are contained. The features encoding the relationship of the term of interest to the retrieved sentences appear to be useful as well.

8.0 FINE-TUNING PRE-TRAINED LANGUAGE MODELS FOR RETRIEVAL OF USEFUL SENTENCES

In this chapter I present experiments analyzing yet another approach to ranking sentences automatically. I explore fine-tuning of pre-trained language models (those released with [67]) for sentence classification and sentence pair classification to tackle the task of ranking the sentences with respect to their utility. Here, a learning algorithm does not start with a hand-crafted feature vector, such as the algorithms analyzed in Chapter 7. Instead, the texts of retrieved sentences (sentence classification) as well as the texts of the terms of interest and their source provisions (sentence pair classification) are utilized directly. A learning algorithm learns the higher-level feature representation as well as the classification function jointly, starting from a model pre-trained in a weakly supervised fashion. This approach has recently proven to work very well on a large number of traditional NLP tasks as evidenced by the leaderboard associated with the General Language Understanding Evaluation (GLUE) benchmark.[1] [120] I provide brief background to the idea of pre-trained language models in Section 8.1. In Section 8.2 I present the experimental setup to assess the effectiveness of the approaches. The results of these experiments are reported and discussed in Section 8.3.

8.1 PRE-TRAINED LANGUAGE MODELS

Analyzing the task, as I have done in Chapter 6, in order to come up with the task specific features (Section 7.1) is time-consuming. An obvious question—the one I have not dealt with so far—is if this was necessary. There is a number of well-established document rep-

[1] gluebenchmark.com/leaderboard

resentations that have been successfully utilized in countless document classification tasks. These include high-dimensional bag-of-words document models, such as TF-IDF weighting of token N-grams, or lower dimensional projections of words onto static word embedding vectors. For completeness, I performed an experiment using such representations. Specifically, I used a Random Forest classifier trained on TF-IDF token N-grams and a Multi Layer Perceptron trained on static word embeddings (GloVe). The study confirmed that such a direct approach does not work very well. In fact, these systems barely outperform the Random baseline (see RF-BoW and MLP-WE in Table 22). The reason behind the poor performance is that these often very useful features do not encode the signal which is needed for ranking sentences per the definition of usefulness implemented in this work. This is, among other things, due to the fact that such representations do not encode the relationship between the document and the query. Consider the following example:

> And, the Act defines "*aural transfer*" as "a transfer containing the human voice at any point between and including the point of origin and the point of reception."

If "aural transfer" is the term of interest, as is the case here, then this sentence would have a 'high value' provided this is not just a verbatim citation from the source provision. However, if, say, "human voice" would be the term of interest the sentence would only have 'certain value.' It should be apparent that the sentence detached from the term of interest (and its source provision) does not provide reliable grounds for determining its value. For completeness, it is important to mention that sentences themselves do carry certain signal. This is evidenced by the performance of BERT snt model described and analyzed later in this chapter. However, the effectiveness is still much lower than what can be achieved if the relationships are taken into account.

In order to establish the relationship between the query and the document using the above mentioned representations, one is limited to techniques such as measuring similarity between the query and the document both projected onto the same space. However, query–document similarity and relevance are not the same. This is especially true in the context of a very specific notion of relevance, such as the one this work deals with. Indeed, the experiments in Chapter 6 show that while the similarity is certainly indicative of sentences' utility, there is much more going on (novelty, topical similarity, context among others). Therefore, a more

sophisticated representation of the relationship is required. One can either hand-craft such a representation (as I did in Chapter 7) or let the learning algorithm do this automatically (this chapter).

It has become common practice in ML for NLP to use language representations pre-trained on some data rich task. For example, static word embeddings, such as those used earlier in this work (e.g., word2vec [77, 76], GloVe [87]) are learned on tasks that encourage co-occurring words to be mapped on similar vectors. Recently, deep neural models pre-trained on language modeling tasks (masked token prediction, next sentence prediction) achieved impressive results on a wide variety of NLP tasks. The notable such models that significantly improved on the then existing state-of-the-art include ELMo [88], OpenAI GPT [94], ULMFit [52], and BERT [23]. Especially, BERT (bidirectional encoder representation from transformers), based on the Transformer architecture from [117], has gained immense popularity. A large number of models using similar architectures have been proposed since then (e.g. RoBERTa [67], ALBERT [63], T5 [95], StructBERT [121]). The core capability of these models, the one which is leveraged in this work, is their fine-tuning on a downstream task. The original model is typically trained on large corpora of general language resources, such as Wikipedia or book corpora, to perform weakly supervised tasks such as masked token prediction or the next sentence prediction. For a downstream task one typically just ads a small layer to the core model that handles, e.g., the classification into four classes (this work). Using a task specific data set, the augmented model is then further trained (fine-tuned) starting from the parameters optimized during the pre-training phase. These models are thus able to perform very well on tasks with little available data, by leveraging the previously learned general understanding of a language.

The models based on the BERT architecture have been successfully used in a variety of IR tasks as well. Several simple applications of BERT to ad hoc document retrieval are presented in [125]. The authors of [70] demonstrate improvement of the state-of-the-art in ad hoc document retrieval by using the BERT's classification vector in existing neural models. Successful applications of BERT for retrieval of short texts such as sentences are presented in [127] and [97]. An end-to-end open source retrieval system utilizing BERT has been released with [126]. Similar to the utilization of source provisions in this work, the authors of [75]

demonstrated the effectiveness of using a query context in a re-ranking component based on BERT. In [86] BERT is fine-tuned on query-retrieved document pairs as is done in this work.

There are examples of successful applications of BERT on legal texts as well. In [16] BERT is evaluated on classification of claim acceptance given judges' arguments. A task of retrieving related case-law similar to a case decision a user provides is tackled in [101]. The authors demonstrate the effectiveness of using BERT for this task while focusing on mitigating the constraint on document length imposed by BERT. In [14] BERT is evaluated as one of the approaches to predict court decision outcome given the facts of a case. BERT has been successfully used for classification of legal areas of Supreme Court judgments. [53] The authors of [93] combine BERT with simple similarity measure to tackle the challenging task of case law entailment. Using this approach they achieved state-of-the-art results on the special task. BERT was also used in learning-to-rank settings, as is done in this work, for retrieval of legal news. [103]

8.2 EXPERIMENTS

The experimental setup is similar to the one used in Chapter 7. I first retrieve all the sentences that contain the term of interest. Using different pre-trained language models fine-tuned on the sentence value classification task, sentences are ranked from the ones predicted as the most valuable to those deemed the least valuable. As baselines I report the performance of a Random system on a large sample of repeated runs (for reference) as well as the BM25 and BM25-c methods from Chapter 6 for statistical testing. For reference, the performance of the Random Forest model trained on the TF-IDF $\{1, 2, 3\}$-gram features and Multi-layer Perceptron model trained on document centroids derived from word embeddings generated by the GloVe model (300 dimensions)[2] from [87], are reported.

As in Chapter 7 the experiments in this chapter are performed on all 42 queries. Again, each method is evaluated in terms of two experimental setups. The first setup is the same 6-fold cross-validation as the one performed in Chapter 7. The second setup focuses on

[2] github.com/RaRe-Technologies/gensim-data

performance of the method on the second and sixth folds that were originally held-out. Here, the need to perform separate experiments with the held-out set may seem less obvious than it did in Chapter 7. The knowledge gained from Chapter 6 is not used to handcraft the list of features. However, the knowledge about the relationship between the term of interest, the source provision, and the retrieved sentence is used in the proposed sentence pair classification models (see below). Since this knowledge was obtained via the analysis conducted in Chapter 6, it is necessary to perform the experiments on the held-out set as well.

In this work I use the RoBERTa (a robustly optimized BERT pretraining approach) described in [67] as the starting point for my experiments.[3] Out of the available models I chose to work with the smaller roberta.base model that has 125 million parameters. This choice was motivated by the ability to iterate the experiments faster when compared to working with roberta.large with 355 million parameters. RoBERTa is using the same architecture as BERT. However, the authors of [67] conducted a replication study of BERT pre-training and found that BERT was significantly undertrained. They used the insights thus gained to propose a better pretraining procedure. Their modifications include longer training with bigger batches and more data, removal of the next sentence prediction objective, training on longer sequences on average (still limited to 512 tokens), and dynamic changing of the masking pattern applied to the training data. [67]

To test the hypotheses S8.1–S8.3 I conduct three experiments (i.e., the sentence classification and the two sentence pair classification experiments). In the first experiment I fine-tune the base roBERTa model on the task of classifying sentences in terms of their value for the interpretation of statutory terms. I apply the fine-tuned model to predict the value of the sentences that were not seen during fine-tuning. By applying softmax to the final prediction layer I obtain the probability distribution over the four possible classes.

$$\sigma(\vec{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{4} e^{z_j}} \text{ where } \vec{z} = (z_1, z_2, z_3, z_4)$$

Here, $\vec{z}$ is a vector of scores for each of the four classes. To obtain a sentence's score I compute an inner product between the per-class probability vector ($\vec{p_i}$) and value weight

[3] github.com/pytorch/fairseq/tree/master/examples/roberta

vector ($\vec{w_i}$) as was done in Section 7.2. The resulting rankings are then scored and compared to the three baselines. Henceforth, this model is referred to as BERT snt.

In the second experiment I fine-tune the base roBERTa model on the sentence pair classification task. Here the model is fed the term of interest as the first sentence and the retrieved sentence as the second one. The task of predicting sentence value is thus recast as predicting the relationship between the term of interest and the retrieved sentence. The goal is still to predict one of the four sentence value labels, i.e., one of the 'no value,' 'potential value,' 'certain value,' or 'high value.' As in the previous experiment, I apply softmax to the final prediction layer to obtain a probability distribution over the classes. Sentences' scores are determined in the same way as in the first experiment. The resulting rankings are scored and compared to the three baselines. Henceforth, this model is referred to as BERT qry2snt.

The third experiment is similar to the second one. Here, I again fine-tune the base roBERTa model on the sentence pair classification task. Unlike in the second experiment, the model is fine-tuned on the source provision as the first sentence and the retrieved sentence as the second one. Therefore, in this experiment the task is understood as prediction of the relationship between the source provision and the retrieved sentence. As in the two previous experiments, softmax is applied to the final prediction layer and the probability distribution over the classes is obtained. The resulting rankings are scored and compared to the three baselines. Henceforth, this model is referred to as BERT sp2snt.

8.3 RESULTS AND DISCUSSION

The results of the experiments described in Section 8.2 are reported in Table 22 (group and overall means) and Figure 35 (box and whisker plots and swarm plots). First, it is clear that the models using shallow features (RF-BoW and MLP-WE) barely outperform the Random baseline. They are not competitive with the methods based on pre-trained language models based on deep architectures. Second, it appears that the methods using fine-tuning of the base roBERTa model are performing better than the two baselines (BM25 and BM25-c).

First, I examine the thesis statement S8.1 by testing the corresponding null hypothesis.

As the first step I use the Friedman test to assess the hypothesis that at least one of the BM25, BM25-c, and BERT snt methods differs from the other 2. The methods are compared in terms of their performance as measured with NDGC@100 on the Overall. I reject the null hypothesis for the cross-validation setup (p-value=.03) but fail to reject it for the held-out set setup (p-value=.14). Hence, I can proceed to test if the BERT snt method is different from the two baselines in the cross-validation experiment. I use the Holm procedure as described in Subsection 5.3. The comparison between BM25 and BERT snt has the lowest p-value (.011). When I adjust the p-value for 2 comparisons, I reject the null (p-value=.022). Since I rejected the null hypothesis I can step down to test the comparison between BM25-c and BERT snt. Here, I cannot reject the null (p-value=.504). I have established that the BERT snt method performs significantly better than BM25 under the cross-validation experimental setup.

Second, I examine the thesis statement S8.2. As the first step I use the Friedman test to assess the hypothesis that at least one of the BM25, BM25-c, and BERT qry2snt methods differs from the other 2. I reject the null hypothesis for the cross-validation setup (p-value=.006) but fail to reject it for the held-out set setup (p-value=.26). Hence, I can proceed to test if the BERT qry2snt method is different from the two baselines in the cross-validation experiment. The comparison between BM25 and BERT qry2snt has the lowest p-value (.006). When I adjust the p-value for 2 comparisons I reject the null (p-value=.012). Since I rejected the null hypothesis I can step down to test the comparison between BM25-c and BERT qry2snt. Here, I cannot reject the null (p-value=.14). Hence, I have established that the BERT qry2snt method performs significantly better than BM25 under the cross-validation experimental setup.

Finally, I assess the thesis statement S8.3. As the first step I use the Friedman test to assess the hypothesis that at least one of the BM25, BM25-c, and BERT sp2snt methods differs from the other 2. I reject the null hypothesis for the cross-validation setup (p-value$< 10^{-7}$) as well as for the held-out set experiment (p-value=.002). Hence, I can proceed to test if the BERT sp2snt is different from the two baselines in both of the experimental setups. The comparison between BM25 and BERT sp2snt has the lowest p-value in the cross-validation ($< 10^{-8}$) as well as in the held-out set experiment (p-value=.0003). After adjusting

Method	SmSp @10	SmSp @100	SmDs @10	SmDs @100	LgSp @10	LgSp @100	LgDs @10	LgDs @100	Overall @10	Overall @100
CROSS-VALIDATION										
Random	.38 ± .10	.67 ± .15	.52 ± .07	.76 ± .09	.25 ± .16	.29 ± .18	.47 ± .09	.48 ± .09	.43 ± .13	.63 ± .21
BM25	.47 ± .13	.74 ± .11	.60 ± .18	.79 ± .11	.44 ± .21	.37 ± .22	.61 ± .17	.56 ± .12	.54 ± .18	.68 ± .20
BM25-c	.48 ± .12	.76 ± .09	.59 ± .17	.80 ± .11	.49 ± .14	.42 ± .17	.63 ± .19	.55 ± .13	.55 ± .16	.70 ± .18
RF-BoW	.47 ± .11	.72 ± .13	.56 ± .16	.77 ± .10	.24 ± .22	.27 ± .20	.59 ± .12	.54 ± .15	.49 ± .19	.65 ± .22
MLP-WE	.40 ± .14	.68 ± .18	.53 ± .15	.75 ± .14	.34 ± .20	.35 ± .18	.66 ± .14	.59 ± .15	.48 ± .18	.65 ± .21
BERT snt	.50 ± .18	.71 ± .17	.61 ± .14	.80 ± .11	.46 ± .24	.47 ± .21	.83 ± .15	.77 ± .12	.59 ± .20	.72 ± .18
BERT qry2snt	.59 ± .23	.76 ± .19	.72 ± .18	.85 ± .10	.64 ± .34	.50 ± .28	.86 ± .25	.77 ± .18	.69 ± .24	.77 ± .20
BERT sp2snt	.57 ± .19	.80 ± .12	.74 ± .15	.87 ± .07	.73 ± .12	.59 ± .18	.89 ± .16	.80 ± .14	.71 ± .19	**.80 ± .14**
HELD-OUT SET										
Random	.34 ± .14	.65 ± .16	.52 ± .05	.75 ± .07	.16 ± .18	.17 ± .17	.47 ± .03	.49 ± .00	.41 ± .16	.60 ± .23
BM25	.41 ± .04	.73 ± .09	.57 ± .20	.77 ± .10	.40 ± .45	.14 ± .07	.56 ± .18	.53 ± .16	.50 ± .20	.63 ± .24
BM25-c	.50 ± .06	.76 ± .05	.55 ± .17	.79 ± .08	.46 ± .28	.23 ± .01	.56 ± .16	.52 ± .13	.52 ± .15	.66 ± .22
RF-BoW	.48 ± .14	.72 ± .13	.59 ± .12	.78 ± .08	.21 ± .23	.20 ± .19	.52 ± .10	.50 ± .05	.49 ± .18	.64 ± .23
MLP-WE	.40 ± .24	.65 ± .29	.47 ± .12	.75 ± .09	.26 ± .28	.27 ± .21	.73 ± .09	.59 ± .13	.46 ± .21	.63 ± .23
BERT snt	.45 ± .13	.66 ± .24	.67 ± .02	.83 ± .04	.39 ± .32	.37 ± .36	.82 ± .16	.75 ± .06	.59 ± .19	.70 ± .22
BERT qry2snt	.41 ± .19	.66 ± .23	.73 ± .10	.86 ± .03	.48 ± .68	.43 ± .53	.98 ± .03	.81 ± .13	.64 ± .30	.73 ± .24
BERT sp2snt	.53 ± .27	.78 ± .17	.67 ± .13	.82 ± .06	.71 ± .23	.48 ± .26	.95 ± .07	.84 ± .02	.68 ± .21	**.77 ± .17**

Table 22: The table shows the results of the experiments with pre-trained language models. The NDGC@10 and NDGC@100 are shown for the small sparse queries (SmSp), small dense queries (SmDs), large sparse queries (LgSp), large dense queries (LgDs), and all of them together (Overall).

the p-values for 2 comparisons I reject the null for both experimental setups ($< 10^{-7}$ and .0007). Since I rejected the null hypothesis I can step down to test the comparison between BM25-c and BERT sp2qry. Here, I also reject the null for both the experimental setups—cross-validation (p-value=.0002) and held-out set (p-value=.012). Hence, I established that BERT sp2snt method performs significantly better than BM25 and BM25-c under the cross-validation as well as the held-out set setup.

The performance of the methods based on the deep pre-trained language models is very promising as can be seen in Figure 35. Even the performance of the naive model that just considers the sentence itself (BERT snt) and is completely oblivious to the query or the source provision shows promise. The performance of BERT snt is especially interesting if one compares it to the performance of RF-BoW and MLP-WE in Table 22 and the performance of the sentence system in 21. From the poor performance of RF-BoW, MLP-WE, and sentence

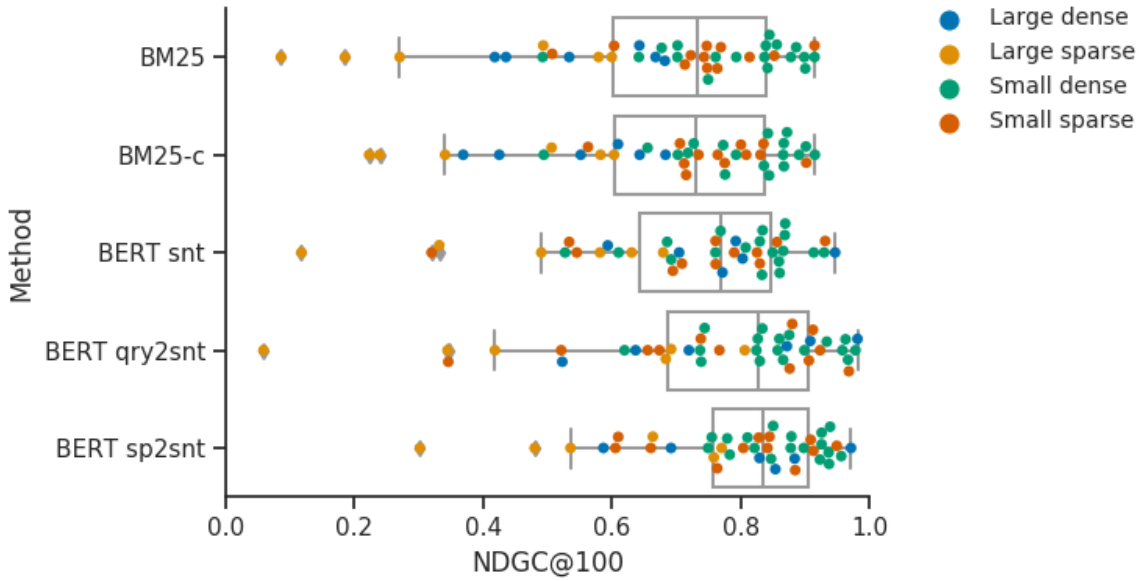

Figure 35: The figure shows scatter plots of the performance on the individual 42 queries measured in terms of NDGC@100. All the three classifiers based on the pre-trained language model appear to perform better than the two baselines.

models (ablation study in Section 7.4), one would be inclined to conclude that there is little signal encoded in the sentence itself. Yet, BERT snt clearly shows that there is some signal afterall. In light of this conclusion, it is interesting to understand what types of sentences are recognized as valuable or not valuable by BERT snt.

First, it is clear that the system correctly recognized that very short pieces of texts that do not form full grammatical sentences typically have very little value. For example, the following sentences have been placed at the bottom of their respective rankings:

i. Communication & Navigation equipment
[term of interest: "navigation equipment;" gold label: 'no value']
ii. B. Non-Disclosure of PreExisting Works
[term of interest: "preexisting work;" gold label: 'no value']
iii. Independent Economic Value
[term of interest: "independent economic value;" gold label: 'no value']

Second, it seems that the system uses features such as the presence of numbering in the sentence, complicated sentence structures, abrupt endings or starts of the sentences, and ref-

erences, to recognize quotations of statutory texts and assign a low value to these sentences.

i. The Electronic Communications Privacy Act of 1986 (ECPA), Pub. L. 99-508, §101(a)(6)(C), 100 Stat. 1848, 1849 (1986), codified, as amended, at 18 U.S.C. §2510(18) (1986), defines "aural transfer" to mean "a transfer containing the human voice at any point between and including the point of origin and the point of reception."
[term of interest: "aural transfer;" gold label: 'high value']

ii. Derives independent economic value, actual or potential, from not being generally known to the public or to other persons who can obtain economic value from its disclosure or use; and [f] (2)
[term of interest: "independent economic value;" gold label: 'no value']

This strategy makes sense in general. The quotation could either be the citation of the source provision ('no value') or a citation of a different provision (high chance of different meaning and hence lower value of a sentence). However, the strategy does not always work. For example, the above shown sentence (i) has been ranked very low but in has 'high value' and should have been ranked high. The 'aural transfer' is a rare example of a term of interest for which there is a statutory definition that is not the source provision in the data set. As a result BERT snt underperforms even the Random baseline on this particular term of interest (NDGC@100 .53 vs .62).

BERT snt also seems to have developed a certain tendency to rank sentences where something is claimed to be something else as having 'high value.'

i. Screen output is considered an audiovisual work that falls within the subject matter of copyright.
[term of interest: "audiovisual work;" gold label: 'high value']

ii. Here, it is perfectly clear that CSS is a technological measure that effectively controls access to plaintiffs copyrighted movies because it requires the application of information or a process, with the authority of the copyright owner, to gain access to those works.
[term of interest: "technological measure;" gold label: 'high value']

iii. Avionics are aircraft radios and navigation equipment.
[term of interest: "navigation equipment;" gold label: 'potential value']

This also appears to be a good strategy that works well many times but not always. For example, the last sentence is just 'potential value' because it uses "navigation equipment" in a different meaning (avionics instead of seafaring).

The above examples demonstrate that the pre-trained deep architecture is capable of detecting very complex features. These are quite difficult to hand-craft using the approach presented in Chapter 7. While it is not difficult to come up with features such as sentence length, as I did, it is far more difficult to come up with features suggesting complicated sentence structures, abrupt endings, or subsumption. It is even more difficult to ensure that

all the relevant phenomena are captured. This work then is a special example showing the effectiveness of the deep language model architectures that have emerged recently.

BERT qry2snt models the relationship between the term of interest and the retrieved sentence. It appears to perform better than the base BERT snt model. This corresponds to what has been shown in Chapter 6 on indicative features as well as in Chapter 7 on feature engineering-based learning-to-rank approaches. BERT qry2snt has access to the same kind of strategies as BERT snt. However, since it is no longer oblivious to the term of interest it can go further. For example, there is a clear trend of ranking very high sentences containing the term of interest surrounded by quotation marks:

i. The first subsection of that provision, entitled "Navigation Equipment," requires tankers to possess global positioning system ("GPS") receivers, as well as two separate radar systems.
ii. An "aural transfer" is "a transfer containing the human voice" at some point in transmission of the communication.
[term of interest: "navigation equipment;" gold label: 'high value']
iii. The following uses of distilled spirits and wine are regarded as "industrial" and therefore will be excluded from any application of the term "nonindustrial use":
[term of interest: "nonindustrial use;" gold label: 'potential value']
iv. We believe the common meaning and general understanding of the term "switchblade knife" is a knife in which the blade extends and is securely locked open upon the pressing of a button or other mechanism.
[term of interest: "switchblade knife;" gold label: 'high value']

This appears to be a viable strategy. Although, there are instances where it does not work perfectly as evidenced by example iii. Here, the sentence is a borderline case between 'potential' and 'certain value' but it is not a 'high value' sentence.

BERT qry2snt appears to have the ability to recognize certain linguistic relationships between the term of interest and other parts of a sentence. The following sentences were not recognized as valuable by BERT snt but they are correctly ranked very high by BERT qry2snt:

i. Airplanes need wings to fly, but that does not mean that all wing designs have independent economic value.
[term of interest: "independent economic value;" gold label: 'high value']
ii. As explained above, the duty titles in this case do not qualify as identifying particulars.
iii. Beer is defined to be fermented liquor made from grain, and in this country mostly from barley.
[term of interest: "identifying particular;" gold label: 'high value']
iv. And "motion pictures" are "audiovisual works consisting of a series of related images which, when shown in succession, impart an impression of motion, together with accompanying sounds, if any."
[term of interest: "audiovisual work;" gold label: 'high value']

All these examples seem to exhibit certain higher level patterns that are intuitively very appealing. Rewriting the above sentences into such patterns could look like this:

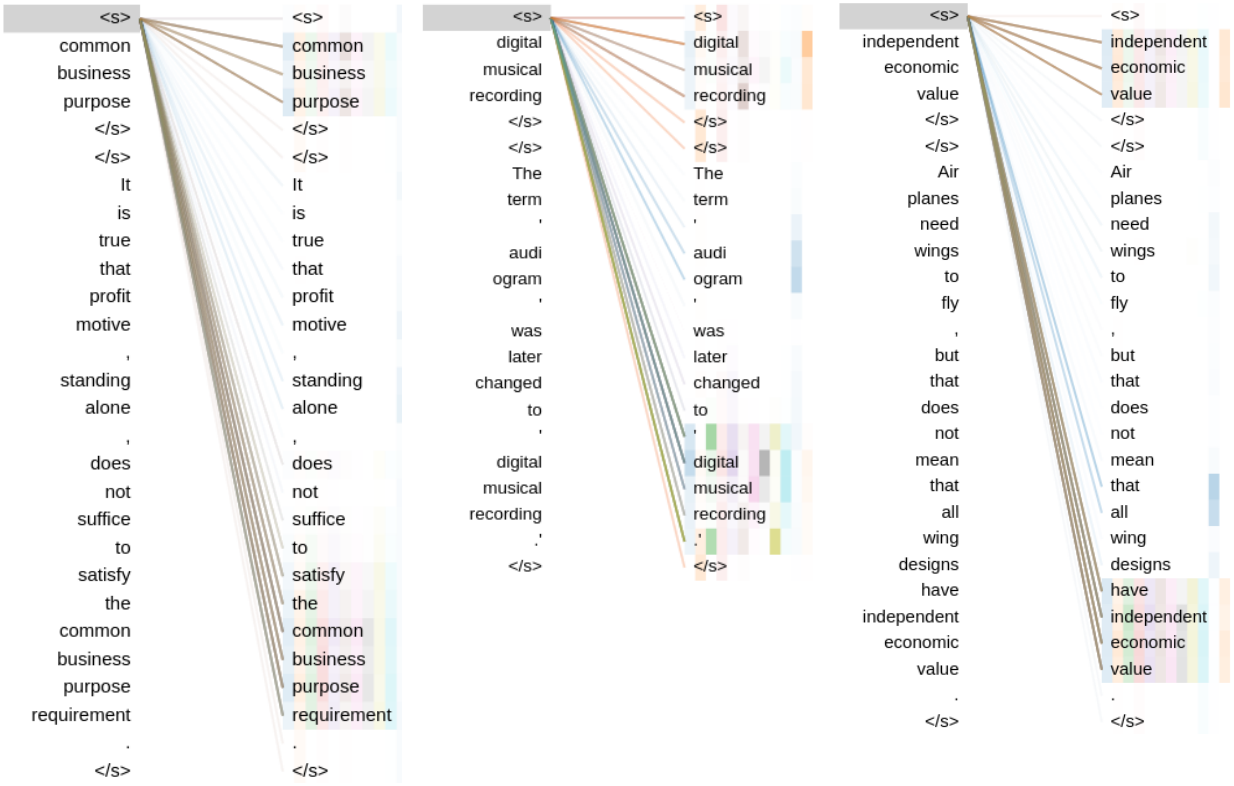

Figure 36: The figure provides some indication of what input elements influence the representation that is being used in the final classification step. The model attends to parts of the sentences that really appear to be suggestive about the higher value of a sentence (i.e., "to satisfy the common business requirement", the quotation marks surrounding the digital musical recording, or "that all . . . have independent economic value").

i. [. . .] NOUN_PHRASE have TERM_OF_INTEREST
ii. [. . .] qualify [. . .] NOUN_PHRASE [. . .] TERM_OF_INTEREST
iii. NOUN_PHRASE is defined to be TERM_OF_INTEREST [. . .]
iv. [. . .] "NOUN_PHRASE" are "TERM_OF_INTEREST [. . .]"

This is corroborated by the inspection of the model weights as applied to several sentences as shown in Figure 36. The visualization was created using the tool published with [118]. As mentioned earlier BERT is based on the Transformer model from [117]. An advantage of using the attention-based model is that it can be interpreted via inspection of the weights assigned to different input elements. As the author of [118] warns one needs to be very

conservative with respect to drawing any conclusions. The three diagrams in Figure 36 show how much attention do the first special tokens pay to the individual words in the three input sequences. Note that the input sequences each consist of a term of interest and a retrieved sentence. The reason why the first special token is interesting is that this token stands for the vector representing the sequence which is then fed into a classifier. Hence, the visualization provides some indication of what influences the representation that is being used in the final classification step. All three examples show that BERT qry2snt establishes the relationship between the term of interest (first part of the sequence) and its mention in the sentence. Additionally, the model attends to parts of the sentences that really appear to be suggestive about the higher value of a sentence (i.e., "to satisfy the common business requirement", the quotation marks surrounding the digital musical recording, or "that all . . . have independent economic value").

Manual assembly of patterns like those shown above is a technique that has been used in the past to tackle tasks quite related to the one I deal with in this work. I am also using certain features described in Section 7.1 that were meant to be a proxy for detection of such patterns (e.g., the minimum distance of the matched query to an explanatory word, or the number of explanatory words on the dependency path between the matched query and a sentence root). As evidenced by the results of the ablation study presented in Section 7.4 this effort yielded very little (if any) success. Coming up with explicit pattern matching templates is very time-consuming and might require domain expertise. In [119] the researchers identified a set of patterns (templates) of indicators for explicit definitions. They then operationalized these patterns with rule-like expressions. [119] Given the complexity of legal texts, the limitations of parsing, and the difficulty of manually constructing the pattern-matching expressions, they explored using ML to generate and test new expressions, but with limited success.

The authors of [32] claim that choosing the best definition sentences requires matching surface patterns associated with definitional sentences, i.e. "lexico-syntactic structures that indicate descriptive content at the sentence level." Since these definitional patterns are not sufficient, "some statistical metrics grounded on the frequencies of terms co-occurring with the definiendum" are employed to weight and rank the sentences," since "highly correlated

words are very likely to express various facets of the definiendum."

Using definitional soft pattern matching, a kind of instance-based learning, and information gleaned from definitional websites, [19] demonstrated that unsupervised learning of soft matching patterns performed better than rule induction or manually-constructed rules, especially when integrating the web information for adjusting ranking weights. Another approach focuses on obtaining patterns, not from word correlations or frequency counts in articles about a particular definiendum, but "across several definitions of the same context or type (e.g., 'is → novelist → romantic')" [32]. There, the focus is on regularities exhibited in lexicalized dependency paths, "sequences of words syntactically and semantically connected within a sentence." These contextual regularities or context models are learned from sample sentences taken from Wikipedia abstracts. N-gram language models (i.e., probability distribution over sequences of words) are constructed for each of the patterns.

Interestingly, in [108] we identified the need for a deeper semantic analysis of the sentences (perhaps, focused on finding typical patterns as in [119]). This was claimed with respect to one of the problems that we observed, but were not able to deal with via methods presented in [108] (roughly corresponding to a subset of methods from Chapter 6). The problem was related to isolated mentions of the term of interest in sentences that clearly have 'high value.' An example is the following sentence from *In re: The Estate of Elvadine Ridgeway, deceased*:

> It is clear [. . .] that the petitioner need not provide identifying particulars (such as account numbers) in the initial petition.

Here, the sentence clearly provides an example of an "identifying particular" (an account number). Yet, since this is the only mention of the term in the whole decision, the sentence is not recognized as valuable. Perhaps a following pattern could inform the model that the sentence should be ranked higher:

> term_of_interest "(such as" _")"

Finally, the BERT sp2snt model that focuses on the relationship between the source provision and the retrieved sentence appears to perform better than the BERT qry2snt model. This may seem somewhat surprising because BERT sp2snt does not have access to the term of interest. On the other hand, it is provided with the full source provision. This

fully confirms what I presented in the ablation study in Section 7.4 related to the systems trained on manually engineered features. There, I showed that the features encoding the relationship between the source provision and the retrieved sentences are by far the most important ones.

While BERT sp2snt appears to lack the ability of BERT qry2snt to detect the useful linguistic patterns attached to the terms of interest, it has the ability to recognize the sentences with 'no value' with a high level of accuracy. For example, the following sentences are ranked very high by BERT qry2snt:

i. In that article, a "wire communication" is defined as "an aural transfer made in whole or in part through the use of facilities for the transmission of communications by the aid of wire, cable, or other like connection between the point of origin and the point of reception."
ii. The semiconductor chip product in turn is defined as: the final or intermediate form of any product–
iii. The term "regular compensation" or "regular military compensation (RMC)" means the total of the following elements that a member of a uniformed service accrues or receives, directly or indirectly, in cash or in kind every payday: basic pay, basic allowance for quarters (including any variable housing allowance or station housing allowance), basic allowance for subsistence; and Federal tax advantage accruing to the aforementioned allowances because they are not subject to Federal income tax.

While these sentences appear to offer a valuable definition of the terms of interest and are thus ranked highly by BERT qry2snt, they merely quote the source provision, and thus have 'no value.' Overall, it appears that with respect to the NDGC scores, it is extremely important to make sure that sentences such as these do not appear at the top positions of the rankings.

8.4 CONCLUSIONS

In this chapter I experimented with fine-tuning a language model based on a deep neural architecture to the task of multi-label classification of sentences according to their usefulness for the interpretation of statutory terms. I experimented with three different approaches: (i) classification of retrieved sentences; (ii) classification of the relationship between the term of interest and the retrieved sentences; and (iii) classification of the relationship between the source provision and the retrieved sentences. I showed that the approach supports learning of a classification function that serves as a basis of a sophisticated ranking function outper-

forming the BM25 and BM25-c baselines (BERT sp2snt). The finding that the relationship between the source provision and the retrieved documents is the most important corresponds to what I showed in Chapter 7.

9.0 DISCUSSION

The preceding chapters present and discuss different approaches to model sentence retrieval for argumentation about the meaning of statutory terms as a sentence ranking task. In this chapter the focus of discussion includes all of the material covered in Chapters 6–8. Here, an overall trend of improving performance as more sophisticated methods are used becomes apparent. First, I explain the selection of the particular methods that are discussed in this chapter. Second, I elaborate on the progress in performance as the focus moves from simpler measures based on similarity or novelty and their combination to more sophisticated learning-to-rank methods based on the feature-engineering approach and pre-trained language models.

Chapter 6 presented a largely manual approach to tackle the sentence retrieval task. There I evaluated a considerable number of traditional text-to-text matching methods, such as similarity measures based on exact token matching (e.g., BM25) or matching based on topical overlap (e.g., LDA), or novelty detection methods. The notable property of these methods is that they are largely unsupervised. The learning is limited to setting a small number of parameters such as k_1, $k3$, and b in the case of BM25. In my work I used the gold labels to optimize these parameters. However, they have an intuitive meaning and since their number is limited, it is often the case that one sets them heuristically based on some insights as to their effects on the rankings. This is exactly what I did when I proposed the two compound methods each of which used several of the traditional measures to create even better rankings. The selection and setting of those constituents was done heuristically based on my intuitions about the task that I had developed during the analysis in Chapter 6. For discussion here I choose to work with the BM25 and BM25-c methods presented in Chapter 6. These methods are very close to what is typically used in a lot of IR systems. Hence, they are very effective baselines that are often not easy to outperform. Furthermore, I include

the BMp+NW+LDA method developed as bundle that brings together the advantages of several such methods.

In Chapter 7 I experimented with a radically different approach. Instead of carefully choosing which of the base measures should go into the compound model and how they should be used I created an extensive list of features. While still using my intuition about what features might be useful, I could afford to be much more liberal and include any feature I thought might be useful. Importantly, I did not need to worry about how exactly a specific feature would be used and how would it combine with the other features. This burden was passed onto the learning algorithm, such as Random Forest or Multi-layer Perceptron. For discussion here I include the RF-PWT which is one of the best (if not the best) performing method from those presented in Chapter 7.

In Chapter 8 yet another approach to sentence ranking was analyzed. There, I explored fine-tuning of pre-trained language models for sentence classification and sentence pair classification to tackle the task of ranking the sentences with respect to their utility. Using this approach the learning algorithm did not start with a hand-crafted feature vector. Instead, the texts of retrieved sentences (sentence classification) as well as the texts of the terms of interest and their source provisions (sentence pair classification) were utilized directly. The learning algorithm learned the higher-level feature representations as well as the classification function jointly, starting from a model pre-trained in a weakly supervised fashion. For discussion here I include the BERT qry2snt and BERT sp2snt models that turned out to be very effective bases for ranking with many interesting properties.

Table 23 and Figure 37 report the performance of the six methods selected as important milestones characterizing this work as well as the performance of the Random system. It is apparent that the use of simple BM25 and BM25-c outperforms the Random baseline. This is because these methods capture the signal which models certain aspects of the computational definition of usefulness implemented in this work (as shown in Chapter 6). Despite their similar performance they benefit from completely different phenomena. Intuitively, BM25 ranks high sentences that contain multiple mentions of the term of interest. In this work the method is optimized in such a way that the documents are not penalized for their length. Hence, the system would often prefer very long sentences. Obviously, such a simple approach

Method	SmSp		SmDs		LgSp		LgDs		Overall	
	@10	@100	@10	@100	@10	@100	@10	@100	@10	@100
Random	.40 ± .07	.69 ± .15	.52 ± .08	.76 ± .11	.29 ± .16	.35 ± .18	.47 ± .11	.47 ± .11	.45 ± .12	.64 ± .20
BM25	.47 ± .13	.74 ± .11	.60 ± .18	.79 ± .11	.44 ± .21	.37 ± .22	.61 ± .17	.56 ± .12	.54 ± .18	.68 ± .20
BM25-c	.48 ± .12	.76 ± .09	.59 ± .17	.80 ± .11	.49 ± .14	.42 ± .17	.63 ± .19	.55 ± .13	.55 ± .16	.70 ± .18
BMp+NW+LDA	.55 ± .11	.78 ± .12	.64 ± .14	.82 ± .10	.58 ± .16	.56 ± .02	.65 ± .23	.62 ± .11	.61 ± .15	.74 ± .14
RF-PWT	.60 ± .16	.81 ± .11	.66 ± .12	.83 ± .10	.71 ± .17	.68 ± .08	.67 ± .10	.64 ± .09	.65 ± .14	.77 ± .12
BERT qry2snt	.59 ± .23	.76 ± .19	.72 ± .18	.85 ± .10	.64 ± .34	.50 ± .28	.86 ± .25	.77 ± .18	.69 ± .24	.77 ± .20
BERT sp2snt	.57 ± .19	.80 ± .12	.74 ± .15	.87 ± .07	.73 ± .12	.59 ± .18	.89 ± .16	.80 ± .14	.71 ± .19	.80 ± .14

Table 23: The table shows the overview results of selected notable methods. The NDGC@10 and NDGC@100 are shown for the small sparse queries (SmSp), small dense queries (SmDs), large sparse queries (LgSp), large dense queries (LgDs), and all of them together (Overall).

works to a certain extent. BM25-c is a combination (linear) of the plain BM25 and another BM25 measure applied to the whole text of a case (i.e., sentence's context). Hence, this system can additionally use the fact of the term of interest appearing many times within the whole text. This is useful because a decision that mentions the term many times is more likely to contain useful sentences than a decision that mentions it just once.

From a number of similarity measuring methods that take sentence's context into account, BM25-p, which uses only a paragraph, instead of the full text of the case, was identified as the most effective. Here, Table 23 and Figure 37 show how a compound method based on this variant of BM25 worked better than the simpler BM25 and BM25-c measures. Apart from benefiting from the similar phenomena as the BM25-c method, the compound system is augmented with two important components. The components model the novelty of the sentence with respect to the source provision (NW) and topical match of the source provision to the full case text (LDA). These work as a proxy for the requirement for a sentence to provide additional information over what is already known from the provision and the requirement to use the term of interest in the same or closely related meaning. The requirements are encoded in the definition of usefulness implemented in this work. Albeit a crude model, the BMp+NW+LDA is a sizable improvement over the Random system that was used as a starting point for the analysis here.

One more improvement in the performance was achieved when I employed a ML algo-

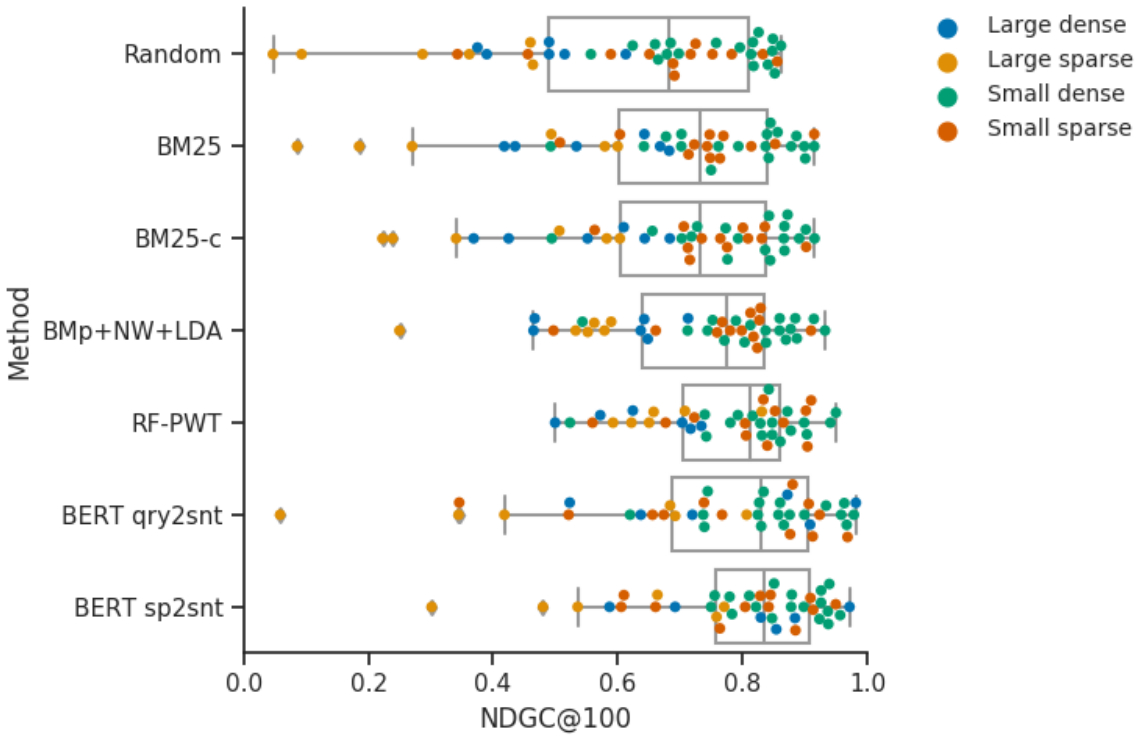

Figure 37: The figure shows scatter plots of the performance on the individual 42 queries measured in terms of NDGC@100. A progression from application of simpler similarity methods towards more complex learning-to-rank systems is shown on a small number of notable methods.

rithm to automatically learn the function that would be a basis for the ranking. Here, the model was given access to 129 features, instead of just 3 as in case of the BMp+NW+LDA. In a sense, the RF-PWT model (Random Forest on the task transformed into pair-wise relevance ranking) is the pinnacle of my efforts presented in Chapters 6 and 7. If one looks at Figure 37 and compares the progression starting from the BM25 method and ending with the RF-PWT he or she should be able to arrive to at least one fairly obvious observation. The large bulk of the improvement subsists in correcting the disastrous performance of the queries on the left tail of the swarm plots. Despite certain improvements happening at the right side as well, these are dwarfed by the events on the left.

The pre-trained language models fine-tuned on the task of sentence pair classification are interesting because they no longer focus on the improvement of the lowest performing queries only. They also bring notable improvement to the queries where the performance had already been decent. This is especially true for the BERT qry2snt model that is completely oblivious to the source provision. Hence, this model cannot address the requirement of "providing additional information" as well as the requirement of "using the term of interest in the same meaning." Indeed, it appears that the model has similar issues with a number of queries that the BM25 method had. Yet, despite these notable issues the overall performance is comparable to (if not better than) the RF-PWT. A huge potential for future work lies in the integration of BERT qry2snt with the BERT sp2snt method that I analyze next.

The BERT sp2snt method uses the source provision instead of the term of interest. Its focus is similar to that of the BMp+NW+LDA and RF-PWT (i.e., the improvement of the most disastrous queries). On closer inspection one sees three large sparse queries that are not handled well by this method. These are 'essential step,' 'identifying particular,' and 'preexisting work.' There are two reasons why a sentence could have 'no value.' It either provides no additional information or it uses the term in a completely different meaning. The three mishandled queries have many sentences that use the term in a different meaning. It appears that BERT sp2snt learned to down-rank the sentences that do not provide additional information quite reliably whereas it completely failed to learn to down-rank the sentences that use the term in a different meaning. The data set may be too small for the method to capture this aspect. Alternatively, a single sentence does not provide as much information as the LDA component from BMp+NW+LDA, that models the phenomenon quite well and uses the text of the whole case. It appears that making the BERT sp2snt method capture the topical similarity would lead to a massive improvement.

Finally, I turn my attention to the analysis of the differences between the performance of the RF-PWT and BERT qry2snt. Since, most of the analysis performed in this work is quantitative, here I offer a qualitative analysis that evaluates the top 10 results for three queries—"navigation equipment," "essential step," and "common business purpose," (see Tables 25, 26, and 24). These queries were selected because all of them are challenging for the manually created BMp+NW+LDA compound system. Hence, they are ideal for study

of the "intelligent" behavior of the two advanced methods I focus on. Overall, the RF-PWT has decent performance on all 3 of the analyzed queries whereas BERT qry2snt appears to completely fail on "essential step" but performs reasonably well on the other two.

Table 24 compares the top 10 results retrieved for "navigation equipment" by the RF-PWT and BERT qry2snt methods. Here, neither of the methods appears to perform very well. At the same time neither of them is terrible. Interestingly, there are only four 'high value' sentences out of the 154 for "navigation equipment." Not a single one of them was missed by BERT qry2sent and all four are in the top 10 results. RF-PWT only managed to recognize two which is still a good outcome. The 'navigation equipment' is a challenging query because it is associated with sentences that use the term of interest in a slightly different, yet related, meaning (aviation or road traffic navigation instead of seafaring). The value of such sentences is demoted per definition of usefulness implemented in this work. As mentioned above BERT qry2sent cannot detect this aspect of the definition because it does not have access to the source provision. Out of the six 'potential value' sentences included in the BERT snt2qry results three use the term in a different meaning. However, if they would use it in the same meaning all three would either have 'certain value' or 'high value.' RF-PWT benefits from its ability to recognize sentences that most likely use the term in a different meaning due to the features that match the source provision to the full text of a case. Therefore, fewer sentences with only the 'potential value' are retrieved. For a complete picture, the lower value classes are very dominant with 119 'potential value' and 15 'no value' instances associated with "navigation equipment" (87%). Hence, both methods clearly show the ability to place the more valuable sentences at the top of the list. Only about 13% of sentences would on average have higher value than 'potential' if the selection were performed randomly whereas BERT qry2sent has 50% and RF-PWT has 80%.

Table 25 compares the top 10 results retrieved for "essential step" by the RF-PWT and BERT qry2snt methods. While the performance of RF-PWT is reasonable, the performance of the BERT qry2snt appears to be a complete disaster. This is the one query where BERT qry2snt completely fails in terms of the NDGC scores (see Figure 37 for the orange dot on the left for BERT qry2snt). The "essential step" query has 2,226 'no value' sentences out of the total number of 2,374, i.e., there is only about a 6.2% chance that the system will randomly

RF-PWT			BERT qry2snt
The first subsection of that provision, entitled Navigation Equipment, requires tankers to possess global positioning system (GPS) receivers, as well as two separate radar systems.	H	H	The first subsection of that provision, entitled "Navigation Equipment," requires tankers to possess global positioning system ("GPS") receivers, as well as two separate radar systems.
The Control Room watch team consists of a: (1) quartermaster of the watch, who is responsible to keep the navigation plot of the vessel on a navigation chart; (2) a helmsman; (3) a navigational electronics technician, who is responsible to maintain the ships inertial navigation equipment; [. . .]	C	H	Penn hired Rhodes to install the BLUE-FIN's remaining electronic navigation equipment, including a Simrad AP50 Plus Autopilot system, which Rhodes did after EMI had already installed the steering.
On May 19, 1977, the M/V FLYING TIGER, also owned by the plaintiff, returned to Shell Key with precision RAY DIST navigation equipment in an effort to locate precisely the object involved in the GOLDEN TIGERs accident.	C	P	This equipment, which is common to most oceanographic research, includes: winches capable of lowering instruments deep into the water; very sophisticated computer installations and precise navigation equipment for accurate maintenance of the vessels position at sea; [. . .]
Penn hired Rhodes to install the BLUE-FINs remaining electronic navigation equipment, including a Simrad AP50 Plus Autopilot system, which Rhodes did after EMI had already installed the steering.	H	P	Avionics are aircraft radios and navigation equipment.
The operation of the Boats calls for the Captains to exercise seamanship and professional judgment, to steer, navigate, and maneuver the Boats, [. . .] to avoid collisions and interference with other vessels, to monitor and operate the Boats radios, to operate the Boats' radars, to operate the Boats' GPS navigation equipment, and to perform towing and related maritime operations.	C	C	The board found the record to show that, since long prior to appellee's date of first use, appellant had used the designation "ARC" as a corporate symbol and as a trademark for airborne communication and navigation equipment, the bulk of which had been sold to the military [. . .]
The ensign, jack, and pennant were kept flying according to regulations; a large quantity of supplies belonging to the Bureaus of Navigation, Equipment, Ordnance, and Steam Engineering were on board; and the regular quarterly returns were made out and transmitted by the plaintiff.	N	H	The vessel was equipped with a type of satellite navigation equipment used for long range navigation more elaborate than the kind of equipment usually used by vessels like the POLARIS.
The pilot house contained expensive satellite navigation equipment and a navigation chart which looked as though it had been used often for charting the area between the Quehara Peninsula in Colombia and the island of Aruba.	C	H	He added "radar" and "SSB Radio" to the list of navigation equipment.
Continental contends that defects in the vessels navigation equipment were concurring contributing causes of the casualty specifically, that the vessels gyrocompass and radio direction finder were malfunctioning, and that certain charts used prior to the grounding were outdated.	C	P	Following is a list of the customer-furnished equipment for each plane involved, and which was installed by Boeing; Galley equipment, radio navigation equipment, including indicating instruments [. . .]
This equipment, which is common to most oceanographic research, includes: winches capable of lowering instruments deep into the water; very sophisticated computer installations and precise navigation equipment for accurate maintenance of the vessels position at sea; sound transducers for research conducted with underwater sound; [. . .]	P	P	For example, car navigation equipment is "difficult to be used portably," is expensive, has limited capacity to store the contents of route guidance, and is unable to provide guidance methods and information to the user beyond conventional route guidance. (155 patent, col. 1:21-42)
Curtis Bay contends, however, that in the present case because the master of the ship was on board during the towing, because some of the navigation equipment aboard the POINT MANATEE was used [. . .] the owners of the POINT MANATEE had sufficient control and direction over the navigation to make pilot Mamo their borrowed servant.	P	P	In 2006, the parties agreed to have Crimson Yachts perform a major overhaul on the BETTY LYN II to include, inter alia, extension of its decks and replacement of its engines, generators, electronics, navigation equipment, plumbing, and wiring.

Table 24: The top 10 results of RF-PWT and BERT qry2snt on "navigation equipment." Sentence value (H: high, C: certain, P: potential, N: no) is reported in the middle.

RF-PWT			BERT qry2snt
That is the import of the phrase "essential step in the utilization of the computer program" that appears in the statute and the phrase "to that extent which will permit its use" that appears in the CONTU Report. does not permit a Nibble purchaser to authorize the defendant to put the programs on a disk for him.	H	N	Indeed, it is not without significance that we have long recognized education as an essential step in providing the disadvantaged with the tools necessary to achieve economic self-sufficiency.
Even though the copy of Vault's program made by Quaid was not used to prevent the copying of the program placed on the PROLOK diskette by one of Vault's customers (which is the purpose of Vault's program), and was, indeed, made for the express purpose of devising a means of defeating its protective function, the copy made by Quaid was "created as an essential step in the utilization" of Vault's program.	H	N	Therefore, in this case, the mailing of the checks was an essential step integral to the completion or fruition of the scheme.
The Sheriffs Department also argued that the copying was an "essential step" under 17 U.S.C. §117(a)(Z), because the hard drive imaging process was a necessary step of installation.	C	N	Therefore, in this case, the mailing of the checks was an essential step integral to the completion or fruition of the scheme.
See [. . .] 17 U.S.C. §117(a) (describing "limitations on exclusive rights" for "computer programs" and providing that infringement does not occur when a copy of a computer program is made either "as an essential step in the utilization of the computer program" or "for archival purposes"); [. . .] ("The term 'literary works" ... includes ... computer programs to the extent that they incorporate authorship in the programmer's expression of original ideas [. . .]	P	N	Although it is true that the goal of the insurers in seeking submission of x rays for use in their review of benefits claims was to minimize costs [. . .] a showing that this goal was actually achieved through the means chosen is not an essential step in establishing that the dentists' attempt to thwart its achievement [. . .]
You may not copy or otherwise reproduce any part of the contents of this package except that you may make one (1) backup copy of the Software and you may load the Software into a computer as an essential step in executing the Software on the computer.	P	N	Similarly, the installation of awning type steel windows and detention screens are an essential step in converting an ordinary room into one for a specific purpose—an escape-proof seclusion room.
One of the grounds for finding that §117 did not apply was the court's conclusion that the permanent copying of the software onto the silicon chips was not an "essential step" in the utilization of the software because the software could be used through RAM without making a permanent copy.	H	N	As to the contentions that the first part of the court's order is too broad, it is inescapably inferred from the evidence that the operations enjoined of unloading, handling, moving, cleaning, cutting, and burning are essential steps in the appellants' business of reconditioning pipe and structural steel.
The question remains whether the changes Titleserv made to its copies of Krause's programs come within §117(a)(l)'s broader concept of an "essential step in the utilization of the computer program in conjunction with a machine."	C	N	The gathering and loading of logs for transport to the sawmill is a necessary and essential step in their processing into lumber.
Even if we were to accept that the Sheriffs Department was the "owner of copies" of RUMBA for §117 purposes, 9 the Sheriffs Department's conduct was not an "essential step," and therefore any error in instruction was harmless.	C	N	Naturally the transportation of logs from woods to mill is an essential step in the manufacture of lumber.
AMI's copying to accumulate a library of the 3090 microcode, or to make copies of the 3090 microcode for reconfigured or split 3090 systems, was not performed as "an essential step" in the use of the 3090 microcode with the 3090 system under §117(1).	H	N	The layout of a highway enlargement and extension by a committee appointed by the common council pursuant to §59 of the Kevised Charter of the City of Bridgeport, is an essential step in the proceedings prescribed by the charter for the establishment of such an improvement [. . .]
The next statutory factor Titleserv must satisfy addresses whether Titleserv's modification of the programs was "an essential step in the utilization of the computer programas] in conjunction with a machine."	C	N	On this point we agree with the Commissioner that precipitation is the essential step in the invention and that "flotation might be replaced by some other method of concentration without destroying the process in issue."

Table 25: The top 10 results of RF-PWT and BERT qry2snt on "essential step." The value of the retrieved sentences (H: high, C: certain, P: potential, N: no) is reported in the middle.

pick a sentence with a value higher than 'no value.' In light of this fact, one can appreciate the behavior of RF-PWT, which menages to retrieve 4 'high value' sentences, 4 'certain value' sentences, 2 'potential value' sentences, and not a single (!) 'no value' sentence.

The "essential step" query is an extreme example where the ability to detect the usage of the term in a different meaning is required. Additionally, it requires the system to understand if additional information is provided over what is known from the source provision. This is another aspect of the definition BERT qry2sent is not able to model. All of the 2,226 'no value' sentences either do not provide additional information or use the term in a completely different meaning. Now, it is true that BERT qry2snt retrieves 10 'no value' sentences. On closer inspection of the results, however, one sees that the behavior of BERT qry2snt is interesting and not a complete disaster. Out of the 10 retrieved sentences, 9 give an example of what is an "essential step" and 1 provides a counter-example, i.e., what is not an "essential step." Per the definition of usefulness implemented in this work such sentences would have 'high value' if the term was used in the same meaning. There are many such sentences among the 2,374 associated with the query. However, there are also many sentences such as the following:

i. The defendant offered no evidence whatever in the proceeding, but to each essential step, as pointed out, filed objections which were overruled by the court.

ii. In each case the statute points out with particularity the essential steps and conditions upon which the right of compulsory appropriation depends, and no conditions other than the law creates can be imposed.

iii. The program operated following seven essential steps.

These sentences mention "essential step" in a way which does not provide much ground for elaboration, and BERT qry2snt has not ranked them highly. Therefore, the selection of the top 10 sentences by BERT qry2sent has some desirable traits despite the result that appears to be a complete disaster.

Table 26 compares the top 10 results retrieved for "common business purpose" by the RF-PWT and BERT qry2snt methods. Performance of both methods is excellent, yet, BERT qry2snt is clearly better. To put things in perspective, "common business purpose" is a rich

RF-PWT		BERT qry2snt	
In all the cases, the fact that the various entities have as a "common business" purpose the profit motive was considered merely incidental and insufficient, standing alone, to support a "common business purpose".	H	H	The unified operation of the restaurants and banquet hall, their similar activities, the centralization of ownership with Timoteo and Maria Manjarrez, the centralization of financial control with Timoteo Manjarrez, the centralization of accounting and administration with Adela Salazar as the general manager of all of the corporate defendants, and the use of Corporate Accounting to administer the payroll are all indicators of common business purpose.
Although a common profit motive alone is not enough to establish a common business purpose, e. g., Hodgson v. University Club Tower, Inc., 466 F.2d 745, 747 (10th Cir. 1972), a common business purpose is present in the instant ease since the motel and restaurant business operations furthered each other "from the standpoint of facilitating the internal operation of the business and from the standpoint of establishing a favorable public image."	H	H	The Fifth Circuit has held that the profit motive is a common business purpose if shared.
By their decisions in the bank cases, Wirtz v. First National Bank and Trust Company, supra and Wirtz v. Savannah Bank and Trust Company of Savannah, supra, found a "common business purpose" between the bank and its office building subsidiary from the following factors:	C	H	It is not believed that the simple objective of making a profit for stockholders can constitute a common business purpose when to achieve this objective one effort engages in banking activities and the other engages in managing an office building and maintaining the same.
Finally, the Secretary argues the businesses operate for a "common business purpose," because they offer each other significant operational advantages and advance the profits of Easton.	C	H	The third requirement—a common business purpose—can be shown through activities that are directed at the "same business objective" or at "similar objectives in which the group has an interest."
Easton also argues that enhancement of profits, without additional interrelationship, does not establish a "common business purpose."	C	H	[A] common business purpose is found in the [banks] operation of the office building to enable the Bank to locate in a desirable downtown area, to provide space for future expansion, to improve the Banks profit position, both from the standpoint of revenue and taxes, and to strengthen the image of the Bank in the public eye.
I would also hold that the cafeteria and the drug store have a common business purpose to the extent that each is part of a "business system" which is directed to a single business objective, that is, the operation of a traditional drug store.	H	H	Further, the operation of the Home as an institution for the care of the aged meets the statutory definition of common business purpose, whether operated for profit or not.
Obviously there is a common business purpose to the extent that both Williams and Fort Payne seek to make a profit and they both serve the traveling public, although presumably the restaurant serves local patrons also.	H	H	The common business purpose is the use of Rockingham Park for horse racing for about two-thirds of each year.
Common Business Purpose "A common business purpose" is also a term used but not defined in the statute or the regulations.	C	H	We find that the grocery store and the slaughterhouse have a common business purpose.
Many of the considerations relevant in determining the existence of related activities are pertinent to determine the existence of a "common business purpose."	C	H	We have held that "the profit motive, standing alone, does not suffice to satisfy the common-business-purpose requirement."
It would be anomalous to treat the owners of commercial buildings as proprietors of individual businesses when they manage the buildings themselves and as participating in a "common business purpose" with other building owners merely because they hire a rental agent who manages other buildings.	H	C	Moreover, the operation of the stores shared a common business purpose.

Table 26: The top 10 results of RF-PWT and BERT qry2snt on "common business purpose." Sentence value (H: high, C: certain, P: potential, N: no) is reported in the middle.

query with 149 out of 880 sentences having 'high value.' Therefore, even the Random system would most likely get one or two 'high value' sentences in the top 10. On the other hand, there are 643 'no value' or 'potential value' sentences (73%). None of those were selected for the top 10 results by either of the methods. The behavior of the two methods differs radically. Upon closer inspection it is not difficult to understand what was driving the preference of the RF-PWT to rank the sentences high. Most of them do mention the term "common business purpose" or its individual constituents multiple times. The remaining sentences contain words such as "profit" or "public" that are semantically very close to the term of interest. Hence, a combination of some token matching feature with some word embedding matching feature is at play here and it works fairly well.

The performance of BERT qry2sent is far more impressive. Nine out of ten sentences it placed on the top have 'high value' and the one 'certain value' sentence appears to be quite a borderline case that could have been labeled as having 'high value' as well. The workings of BERT qry2sent on this query are clearly driven by recognizing the kind of patterns that was discussed in Section 8.3. To be clear, I am not claiming the system uses exact matching patterns to make its decisions. However, the presence of the patterns clearly suggests that the system learned to prefer such sentences.

Overall, the results are very promising. The RF-PWT and BERT sp2snt methods clearly show that it is possible to model the provision of additional information over what is already known from the source provision very accurately. The RF-PWT system demonstrates the same with respect to the term of interest being used in a different meaning. While unable to capture these two important aspects of the definition of usefulness, the BERT qry2snt method exhibits a remarkable ability to select the sentences where the term is mentioned in such a way that the sentences are useful for argumentation about the meaning of the term. This clearly suggests that the integration of these methods will result in a system superior to those presented here. In the short term, one would see a sizeable improvement of BERT qry2sent if it is augmented with the NW+LDA components of the compound system described in Section 6.6 or if the class scores for each sentence generated by BERT qry2snt were plugged as features into RF-PWT. However, from the longer term perspective it appears to make more sense to augment the data set and ensure BERT sp2snt develops the ability to

detect the difference in the meaning between the term of interest and the term used in the sentence. The integration of BERT qry2snt and the augmented BERT sp2snt would most likely lead to a significant upward jump in performance. And since, the performance of the systems I have already evaluated here is very reasonable such a future system has a lot of appeal.

10.0 CONCLUSIONS

10.1 EVALUATION OF THESIS STATEMENTS

In this work I have provided empirical evidence for the proposition that given a statutory provision, a phrase in that provision, and a database of case law, a computer system can autonomously rank the sentences retrieved from the case law in terms of how useful they are for argumentation about the meaning of the phrase. I have shown several approaches leading to systems that are capable of ranking the retrieved sentences with an effectiveness which is significantly better than well-established and often quite difficult-to-beat baselines based on the BM25 measure.

I have identified a number of features that are indicative of sentences' usefulness. Specifically, I have shown that a similarity between the phrase and a retrieved sentence as measured with BM25 or TF-ISF is a statistically significant indicator when compared to random ordering (thesis statement S1.1). When using similarity measures based on word embeddings that consider the relationship between the phrase terms and all the terms in a retrieved sentence I observed a clear trend of their being indicative of usefulness. However, I could not establish statistical significance (thesis statement S1.2). Combining the two types of methods, i.e., those considering the query terms only with those considering all the terms in a sentence for the purpose of similarity assessment, did not appear to be more indicative of usefulness than using the BM25 or TF-ISF methods only (thesis statement S1.3). Despite seeing a trend of being even more indicative, I was not able to show statistically significant improvement for methods that consider not only the sentences themselves but also their contexts. This holds for the methods matching the query terms only (thesis statement S2.1), the word embeddings-based methods matching the whole sentences and contexts (thesis statement

S2.2), as well as their combinations (thesis statement S2.3).

I have established that novelty of a sentence with respect to the statutory provision the phrase comes from is indicative of sentence's relative usefulness. I have shown how filtering of the least novel sentences as measured with NWrds-f, NWrdsW-f, and NWrdsR-f is a statistically significant indicator with respect to random ordering (thesis statement S3.1). I have also demonstrated that using novelty detection methods as full-blown ranking measures does not provide a stronger indication beyond the one coming from their being used just for filtering (thesis statement S3.2). I reached the same conclusion with respect to topical similarity. The fastt-f^{SIF}, DESM-f$_{I\times O}$, and LDA-f methods are statistically significant indicators as compared to random ordering when used for filtering of the topically least-related documents (thesis statement S4.1). Use of these methods on all of the documents does not appear to provide a stronger indication (S4.2).

Quite surprisingly it turned out that the membership of a sentence in a specific functional part of a court decision (e.g., background or analysis) is not indicative of the sentence's usefulness (thesis statement S5.).

I have shown that the features describe different aspects of the definition of usefulness, which means that several features combined provide an even stronger indication. I have demonstrated that a method composed of BM25-p, NWrds-f, and LDA-f is a significantly stronger indicator than two of its constituents (NWrds-f and LDA-f). With respect to BM25-p, I could not establish statistical significance despite coming quite close. The trend is clear (thesis statement S6.1). I observed visible improvement in case of a compound model based on the GloVe-r method as well. However, I could not establish statistical significance (thesis statement S6.2).

I demonstrated that given the set of features indicative of sentences' usefulness, the task of retrieving sentences can be successfully tackled as a learning-to-rank problem. This means that the features enable autonomous learning of a ranking function, which is a reasonable model of sentences' utility for argumentation about the meaning of statutory terms. Specifically, using the set of features put together on the basis of investigating thesis statements S1.–S6., I showed that a sentence classification or regression system for predicting the value of a sentence could be trained as a reasonable model for ranking the sentences according to

their utility. I established that SVM and RF methods significantly outperform the BM25 and BM25-c baselines (thesis statement S7.1). The other ML methods, i.e. LinReg, AdaBoost, LogReg, and MLP, performed visibly better than the baselines as well. However, a statistical significance could not be shown for these. Casting the task into ordinal classification (thesis statement S7.2) or pair-wise relevance classification (thesis statement S7.3) did not appear to lead to visible improvements over the multi-class classification methods.

I showed that pre-trained language models based on deep neural network architectures can be fine-tuned for the special task of retrieving sentences for argumentation about the meaning of statutory terms. This leads to a learning-to-rank system that does not require hand-crafted features. Specifically, I established that predictions of the roBERTa base model fine-tuned on the task of sentence classification in terms of their usefulness (BERT snt) can be utilized as a basis for sophisticated ranking, significantly outperforming the BM25 baseline and visibly outperforming the BM25-c baseline (thesis statement S8.1). Furthermore, I demonstrated that predictions of the roBERTa base model, fine-tuned on the task of sentence pair classification between a term of interest and a sentence with respect to usefulness (BERT qry2snt), can form a basis for sophisticated ranking significantly outperforming the BM25 baseline and visibly outperforming the BM25-c baseline (thesis statement S8.2). Finally, I established that predictions of the roBERTa base model, fine-tuned on the task of sentence pair classification between a source provision and a sentence with respect to usefulness (BERT sp2snt), can be utilized as a basis for sophisticated ranking, significantly outperforming the BM25 and BM25-c baselines (thesis statement S8.3).

10.2 CONTRIBUTIONS EVALUATION

In the Introduction to this work I claimed that it makes a number of contributions to the areas of legal IR and legal text analytics. In this section I would like to consider the contributions in light of the results and outcomes of the work. This work proposed a novel task of discovering sentences for argumentation about the meaning of statutory terms. The task was framed as a sentence ranking problem where the goal was to rank the sentences

from the most to the least valuable ones. Proposing a new task has many pitfalls. As one works through the details of putting together a data set or performing the experiments, one may realize that the task is perhaps not as interesting as it seemed or that there are some problems that were not apparent at the beginning. The fact that now the task is supported by a data set as well as a number of different methods applied to that data set is a great outcome. The fact that there is plenty of space for further improvements suggests that the task will continue to be interesting beyond this work.

To support the experiments presented in this work, the statutory interpretation data set was assembled. The data set comprises 42 queries (i.e., statutory phrases) each of which is associated with a number of sentences retrieved from a sizable corpus of United States case law. Each of the 26,959 sentences was seen by at least three human annotators (14 annotators in total). The data set will be released to the public. Despite certain difficulties I identified during the annotation process and worked hard to remedy, the data set supported many of the experiments presented in this work. Despite its limited size it allowed me to show statistical significance in many of those experiments.

I have been able to show that the task of discovering sentences for argumentation about the meaning of statutory terms is related to document relevance ranking, novelty detection, as well as topic modeling. By doing this I was able to position this novel task in the context of a larger body of work on ad hoc document retrieval. By doing this I hope that I succeeded in demonstrating some appealing applications of those methods.

Based on the detailed task analysis I assembled a list of 129 descriptive features that model the retrieved sentences, their relationships to the statutory phrase as well as the provision of law it came from. These features are domain specific adaptations of similar lists from areas such as general web search. I showed that the proposed feature set could be successfully utilized in learning-to-rank settings by demonstrating how a number of algorithms learn to rank the sentences with a very reasonable quality. By doing this I confirmed the viability of the classical feature-engineering based learning-to-rank approach which has recently been overshadowed by the rise of the deep learning approaches.

Finally, I have shown a possible avenue for future work by establishing that a deep representation appears to be capable of capturing the useful signal for this task, potentially

eliminating the need for hand-crafted features. The whole work then presents a compelling demonstration of the effectiveness of methods based on deep representations in a specific task from a specific domain. This is important because advances in general NLP and ML do not always fully transfer to specialized domains such as legal texts.

10.3 LIMITATIONS AND FUTURE WORK

In order to operationalize the task of discovering sentences for argumentation about meaning of statutory terms, I made a simplifying assumption about the set of relevant sentences. Specifically, only those sentences that have at least one exact mention of the term of interest are considered potentially relevant in this work. In reality, sentences that do not mention the term explicitly could be useful as well. This is a clear limitation of this work. Extending the work so that it would take such sentences into account is a challenging task that I leave for future work.

Due to limited resources it was not possible to include every potential term of interest into the data set. Specifically, I limited the terms to those that returned fewer than 5,000 results (sentences). Extending the data set in general, but especially in such a way that some of the terms with larger results set would be included, is a necessary step for this work to continue successfully. A related phenomenon is the task definition expressed in the annotation guidelines and the quality of the annotation produced by the hired student annotators. Before any future extension of the data set, it is important to address the problems I identified in Section 4.7.

In this work I assumed that the best length for a passage elaborating on the meaning of a statutory term is a sentence. This is not always true. Sometimes the best passage could be just a part of a sentence. Other times it could be multiple sentences or even a whole paragraph. Accounting for this phenomenon would be a very interesting and challenging task that is left for future work.

In this work I retrieve sentences from the opinions of the courts. Court decisions apply statutory provisions to specific cases. To apply a provision correctly a judge usually needs

to clarify the meaning of one or more terms. This makes court decisions an ideal source of sentences that possibly interpret statutory terms. Legislative history and legal commentaries tentatively appear to be promising sources as well. Investigating the usefulness of these types of documents is left for future work.

Here, I have limited the task to ordering the sentences from the most valuable ones to the least. In practice one would also want to account for a possible redundancy of sentences. This means preventing a situation when a user is again offered the exact same or similar sentence at a high place in the ranking. This important step is left for future work.

In this work I only perform an intrinsic evaluation of the performance of the various methods on the task of ranking the retrieved sentences. However, this task is only useful in the context of the larger goal of interpreting statutory terms. It remains to be shown that the improvements in the methods presented here translate into the improved quality or effectiveness of the downstream task. Such an evaluation would at a minimum require human subjects to confirm the utility of the retrieved sentences and their rankings. Ideally, a study showing an improvement in quality or speed of drafting of legal documents (e.g., memos) focused on statutory interpretation when the proposed system is used would be required.

Recently, it has become clear that an effective use of ML and NLP requires certain level of trust and transparency. Essentially, it is important for a user to understand what stands behind the decisions the model is making. In the context of my work, it would be preferable if I could augment the list of sentences with some kind of visual clues that inform the user as to why the system believes a sentence is useful for argumentation about the meaning of a statutory term. In this work the closest I have come to addressing this issue was in the ablation study presented in Section 7.4 and the inspection of the attention mechanism of the BERT qry2snt model in Section 8.3. However, these are inadequate as fully-fledged explanations. Hence, this is another important area left for future work.

Finally, I have only scratched the surface when it comes to the use of pre-trained language models based on deep neural architectures to support the task in this work. There is plenty of space for future work in this respect. A most natural next step would be a model that takes into account the relationships among all three of the basic constituents, i.e., the term

of interest, the source provision, and the retrieved sentences (as explained in Chapter 9). Furthermore, pre-training the model on some weakly supervised task (e.g., masked token prediction) directly on legal texts (e.g., the court decisions) would likely yield interesting results, as well. Integration of the deep learning models with the classical models based on feature engineering would almost certainly lead to a system that outperforms both.

APPENDIX A

INSTRUCTIONS FOR ANNOTATION OF SENTENCES WITH RESPECT TO THEIR GENERAL INTERPRETIVE VALUE FOR A SELECTED PHRASE

Each of the three following sheets contains an annotation task (i.e., the "independent economic value," "identifying particular," and "common business purpose" sheets).

At the top of each sheet there is a cell with a light yellow background that contains an abbreviated text of a single statutory provision. We will call this provision the source provision. In the source provision a short phrase (a couple of words) is printed in blue. This is the phrase of interest, that is, the phrase in whose meaning we are interested.

Below the source provision there is a list of sentences retrieved from the Court Listener web service. These sentences come from the top 20 documents responsive to the query in the form of the phrase of interest in double quotes (e.g., "identifying particular") for 120 Federal jurisdictions as retrieved on February 15, 2016. The sentences are grouped according to the decisions they come from and they are listed in the order in which they appear in the decisions. The annotator's task is to evaluate each sentence in terms of its usefulness for interpretation/explanation of the phrase of interest beyond what is already known from just reading the source provision.

Specifically, the task of the annotator is to decide in which of the following four categories the sentence referring to the phrase of interest belongs:

1) **high value** – This label is reserved for sentences the goal of which is to elaborate on the

meaning of the phrase of interest.

2) **certain value** – An annotator should select this label if the goal of the sentence is not to elaborate on the meaning of the phrase of interest but the sentence still provides grounds to draw some (even modest or quite vague) conclusions about the meaning of the phrase of interest.
3) **potential value** – This label is appropriate if the sentence does not appear to be useful for elaboration on the meaning of the phrase of interest but the sentence provides some additional information (even quite marginal) over what is known from the source provision.
4) **no value** – This label should be selected if the sentence does not provide any additional information over what is already known from the source provision.

For the sake of clarity let us give a couple of examples using the following artificial source provision, coming from, say, New York state law, and the phrase of interest (printed in blue):

"No vehicles are allowed in the park."

The sentences that directly elaborate on the meaning of the "vehicle" belong to the "high value" category.

1. Any mechanical device used for transportation of people or goods is a vehicle.
2. A vehicle usually has wheels, engine and controls.

The sentences that assign or contrast the phrase of interest to some other phrase also belong to the "high value" category.

3. A car is a vehicle.
4. Not every vehicle is a man-made object.

The sentences that can be used to elaborate on the meaning of the "vehicle" but do not directly elaborate on the meaning themselves belong to the "certain value" category.

5. Today I took my horse for a ride in that park where no vehicles are allowed.
6. The main reason why no vehicles are allowed in that park is to secure a tranquil environment there.

The sentences that do not seem to be useful for elaboration on the meaning of the "vehicle" but at the same time provide additional information over what is known from the source provision belong to the "potential value" category.

7. The park where no vehicles are allowed was closed during the last month.
8. The courts often need to analyze the provision stating that "No vehicles are allowed in the park."
9. The Maryland law also provides that "No vehicles are allowed in the park."

If the sentence does not provide any additional information over what is already known from the source provision it belongs to the "no value" category.

10. The provision states that: "No vehicles are allowed in the park."
11. A vehicle is forbidden from entering the park.

Finally, there are three possible scenarios that fall between the cracks of the above categorization. If they occur it is necessary to reassign the sentence to a lower label than would otherwise be assigned to the sentence according to the above rules.

These scenarios include:

a) The sentence uses the phrase of interest from a statutory provision or case law from a different jurisdiction.
b) The sentence is attributed to a person who has a personal interest in interpreting the phrase of interest in a certain way.
c) The phrase of interest in the source provision and the phrase of interest in the sentence have different meanings.

If the sentence uses the phrase of interest from a statutory provision or case law from a different jurisdiction its usefulness should be discounted. If using the standard rules the sentence is assigned to the "potential value" or the "no value" category it is not necessary to do the discounting. Such is the case of the example 9. However, if the sentence is assigned to the "high value" or the "certain value" category it should be re-assigned to a category one step lower.

12. The Indiana law states that: "Vehicle is anything which serves as a means of transport." [Although, this sentence would normally belong to the "high value" category it should be assigned to the "certain value" category to take into account that it describes the state of the affairs in a different jurisdiction.]
13. [Continuation of example 12] Therefore a car is a vehicle. [ditto]

If the sentence is attributed to a person who has a personal interest in interpreting the phrase of interest in a certain way its usefulness should be discounted. If using the standard rules the sentence is assigned to the "potential value" or the "no value" category it is not necessary to do the discounting. If the sentence is assigned to the "high value" or the "certain value" category it should be re-assigned to a category one step lower.

14. The defendant claimed he did not break the rule since roller skates cannot be considered a vehicle. [Although, this sentence would normally belong to the "high value" category it should be assigned to the "certain value" category to take into account that the defendant has a personal interest in interpreting the "vehicle" in this way.]

If the phrase of interest in the source provision and the phrase of interest in the sentence have different meanings the usefulness of the sentence should be discounted. If the meanings are significantly different then the sentence should be labeled as "no value".

15. A body is a vehicle for a soul. [Although, this sentence would normally belong to the "high value" category it should be assigned to the "no value" category to take into account that it uses the "vehicle" in a significantly different meaning.]

If the meanings are different but strongly related the sentence should be re-assigned to a category one step lower if it would normally be labeled with the "high value" or the "certain value" category. If it would normally be labeled with the "potential value" or the "no value" category the usefulness of the sentence should not be discounted.

16. Any autonomous vehicle is subject to the approval of the executive committee. [Although, this sentence would normally belong to the "certain value" category it should be assigned to the "potential value" category to take into account that the "vehicle" is used in a slightly different meaning in the sentence.]

It may happen that the annotator encounters a sentence that does not clearly belong to any one of the four categories. If the annotator is deciding between the two adjacent categories (e.g., between the "potential value" and the "certain value" categories) he should assign the sentence with the one he intuitively feels as more appropriate. However, if the doubt is more serious and the annotator is not deciding merely between the two adjacent categories, the sentence should be flagged as problematic and it should be explained why the annotator thinks it does not belong to any one of the four categories.

APPENDIX B

ANNOTATION GUIDELINES FOR EVALUATING SENTENCES FOR ARGUMENTATION ABOUT THE MEANING OF STATUTORY AND REGULATORY TERMS

B.1 BACKGROUND

Statutory and regulatory provisions are difficult to understand because the rules they express must account for diverse situations, even those not yet encountered. This means the provisions need to be abstract and general. In order to achieve the required generality legislators use vague open textured terms, abstract standards, principles, and values. In order to use such rules successfully it is necessary to map the general norms onto specific factual circumstances. This may often prove to be a considerable challenge.

When the application of a general rule is not straightforward a lawyer must present arguments as to why a provision should be applied in a particular way. In doing so the lawyer must often *defend a specific account of the meaning* of one or more terms. The persuasiveness and validity of a complex argument may hinge on a particular account of the meaning. Argumentation about the meaning of a term may even be the crux of an overall argument.

A thorough analysis of *the past treatment* of the term of interest is foundational to formation of an adequate argument. It is of crucial importance that the proposed account of the meaning of the term can withstand a scrutiny of the available evidence. This evidence consists of past mentions and uses of the term in sentences from documents such as court

decisions, legislative histories, or journal articles.

Consider the following (abridged) excerpt from 29 U.S. Code 203:

> "Enterprise" means the related activities performed [...] by any person or persons for a common business purpose [...]

A lawyer who would like to argue that two restaurants located in different parts of a city owned by a single person do not constitute an enterprise may, e.g., argue that they cannot be considered related activities or that their operation is not performed for a common business purpose. In doing so he would likely refer to the past mentions of the term such as these:

> The fact of common ownership of the two businesses clearly is not sufficient to establish a common business purpose.
>
> The profit motive is a common business purpose if shared.

It should be clear that not all of the sentences are created equal. Some sentences are more useful for the argumentation about the meaning of the term than others. Contrast the two sentences listed above to the following less useful sentences:

> Because the activities of the two businesses are not related and there is no common business purpose, the question of common control is not determinative.
>
> The defendants weakly challenge the common business purpose conclusion.

The ability to sift through large amounts of legal documents and distill the content, that could be subsequently used in argumentation about the meaning of a term, is an important part of any lawyer's skill set. Yet, acquiring this ability requires significant effort. A lot needs to be considered before one understands the value of content that uses the term of interest. These include answering questions such as these:

> Does a sentence provide additional information to what is already known from the statutory provision?
>
> Does the sentence content provide solid grounds for understanding some useful facets of the meaning of the term of interest?
>
> Is the meaning of the term used in the sentence the same as the meaning of the term of interest?

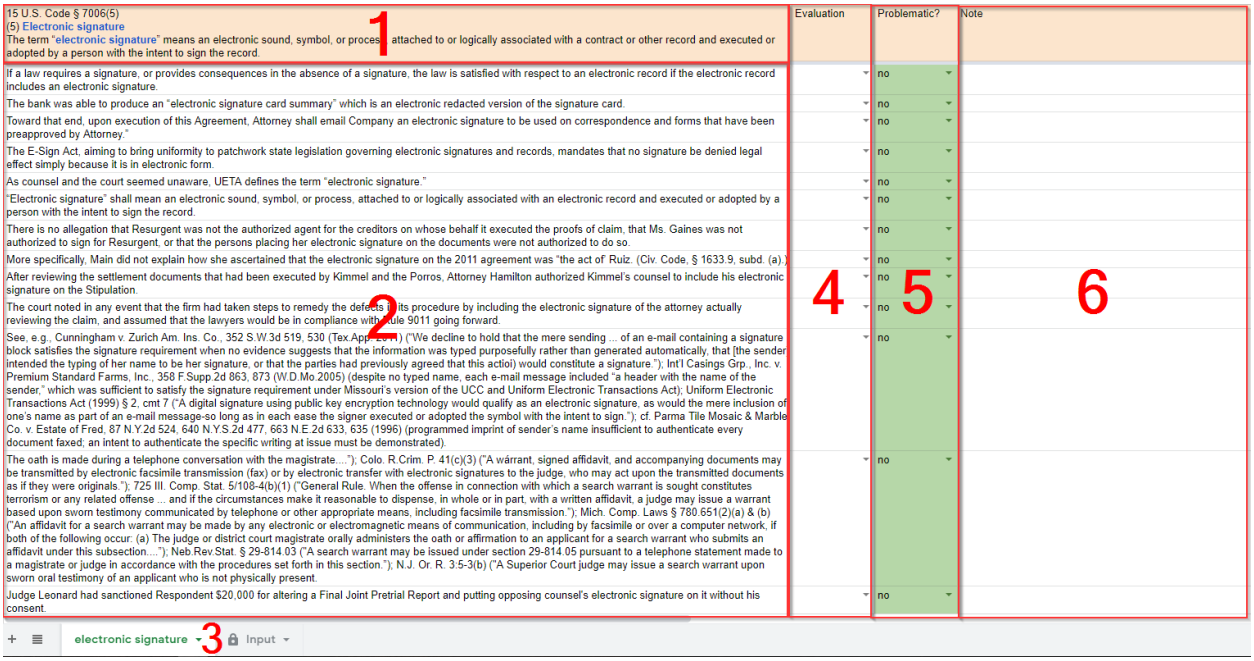

Figure 38: The annotation environment showing the source provision (1), sentence list (2), task list (3), evaluation drop-downs (4), "Problematic?" flag (5), and the note field (6).

B.2 ANNOTATION TASK AND ENVIRONMENT

Each student annotator has access to multiple sheets each of which contains a different annotation task (see 3 in Figure 38). At the top of each sheet there is a cell with a light yellow background that contains a text of a single statutory provision, the source provision (see 1 in Figure 38). In the source provision a short phrase (a couple of words) is printed in blue. This is the phrase of interest, that is, the phrase in whose meaning we are interested.

Below the source provision there is a list of sentences retrieved from U.S. case law (see 2 in Figure 38). These sentences come from the case decisions responsive to a query in the form of the phrase of interest. The sentences are listed in a random order. The annotator's task is to evaluate each sentence in terms of its usefulness for argumentation about the meaning of the phrase of interest. Following the rules described in Section B.3 (below) the annotator should label each sentence with one of the following categories:

1. **High value** - This label is reserved for sentences that explicitly elaborate on the meaning of the phrase of interest.
2. **Certain value** - An annotator should select this label if the sentence does not explicitly elaborate on the meaning of the phrase of interest, yet the sentence still provides grounds to draw some (even modest or quite vague) conclusions about the meaning of the phrase of interest.
3. **Potential value** - This label is appropriate if the sentence does not appear to be useful for elaboration of the meaning of the phrase of interest but the sentence provides some additional information (even quite marginal) over what is known from the source provision.
4. **No value** - This label should be selected if the sentence does not provide any additional useful information over what is already known from the source provision.

Students can select the category for each sentence from a drop-down list to the right of each sentence (see 4 in Figure 38). If the annotator believes that a sentence cannot be evaluated he may indicate so by setting the "Problematic?" flag to "yes" (see 5 in Figure 38). The annotator may optionally use the "Note" field to add a comment (see 6 in Figure 38).

The task of the student annotator is to evaluate all the sentences in all the sheets. Once this is done the annotator should contact jas438@pitt.edu and ask to be assigned additional sentences. We expect each annotator to evaluate at least 1,000 sentences but we expect an average annotator to evaluate about 5,000 sentences. There is no upper limit on how many sentences an annotator may evaluate. However, we would not assign additional sentences to an annotator whose performance turns out to be unsatisfactory.

B.3 RULES FOR SENTENCE EVALUATION

The annotator should evaluate each sentence using the procedure described in this section. Figure 39 provides a flow chart to guide annotators in asking the questions (and indicates which subsection provides more information.) The details about how to answer each question

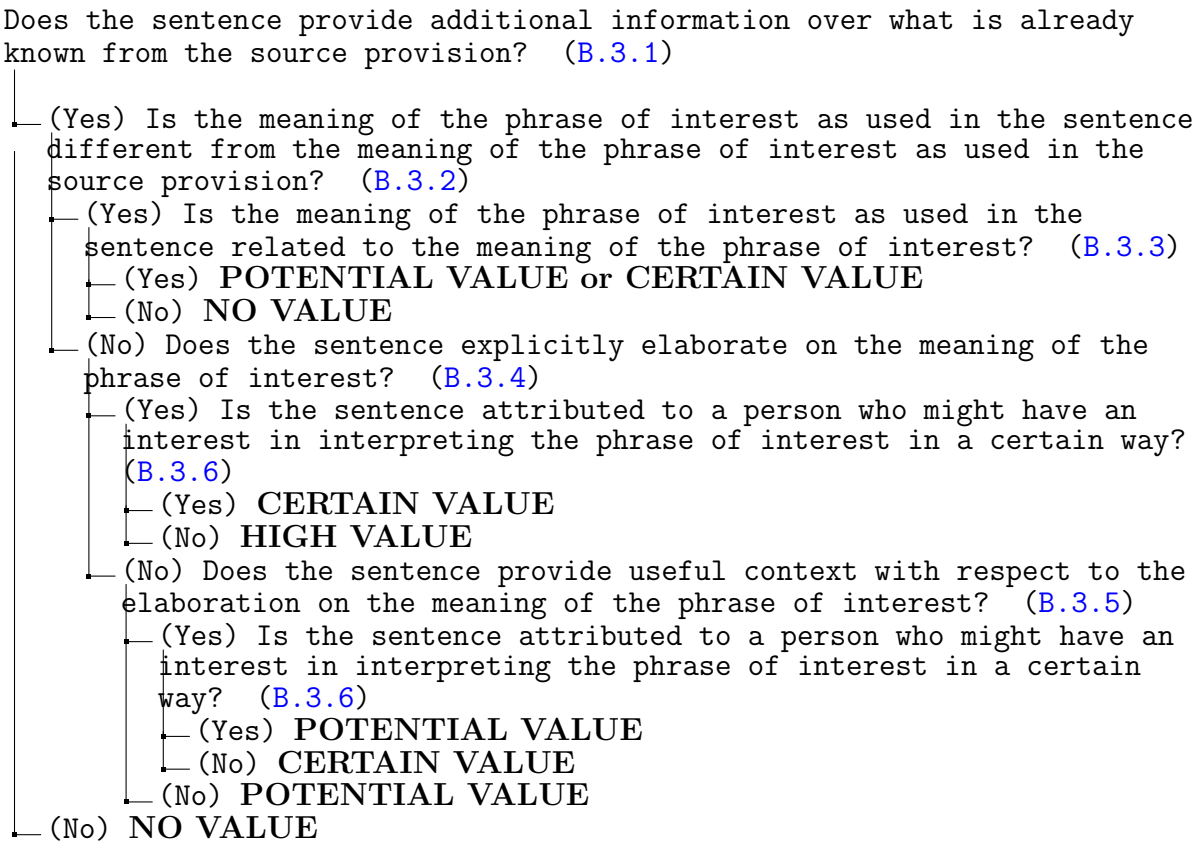

Figure 39: A simplified guidance for classification of the sentences.

are presented in Subsections B.3.1–B.3.6. Annotators should ask the questions outlined in Figure 39 and assign a label based on the answers to those questions.

B.3.1 Does the sentence provide additional information over what is already known from the source provision?

Sentences that contain only verbatim citations or paraphrases of the statutory provision are typical examples of sentences that do not provide additional information. Other examples could be headings or references having no meaningful content beyond the phrase of interest. Consider the following source provision:

> No vehicles are allowed in the park.

The following four examples are sentences that do not provide additional information:

> The provision states that: "No vehicles are allowed in the park."
>
> A vehicle is forbidden from entering the park.
>
> Vehicle
>
> Motor Vehicles Inc. v. Jane Doe

B.3.2 Is the meaning of the phrase of interest as used in the sentence different from the meaning of the phrase of interest as used in the source provision?

It may happen that the phrase of interest from the source provision has a different meaning than the phrase used in a sentence. Again, consider the following provision:

> No vehicles are allowed in the park.

In the following sentence the term *vehicle* is used in a different meaning:

> Any autonomous vehicle is subject to the approval of the executive committee.

Typical examples would involve the same terms or phrases from different legal domains (e.g., "independent economic value" might have a different meaning in the context of copyright law as compared to trade secret protection). Also, where there is a different statute that uses what appears to be the same term, one cannot automatically assume that the terms in both statutes mean the same thing even if the regulatory domains are similar. There are ways to argue that they mean the same thing, but also ways to argue that the terms mean something different. The decision in such cases is left to the annotator's discretion.

B.3.3 Is the meaning of the phrase of interest as used in the sentence related to the meaning of the phrase of interest?

When there is a difference in the meaning of the phrases (per Question B.3.2) it remains to be determined if the difference is as significant as making the phrases effectively unrelated. Again, consider the following provision:

No vehicles are allowed in the park.

In the following sentence the term *vehicle* is used in a related (yet, slightly different meaning):

Any autonomous vehicle is subject to the approval of the executive committee.

Whereas in the sentence below the meaning is very different:

A body is a vehicle for a soul.

The decision as to the degree of the difference in such cases is left to the annotator's discretion. The rule of thumb is if understanding of one of the phrases helps in understanding the other then they are related. If not then the difference is significant. In case the terms are related the sentence is to be assigned with a value one step lower than in case of the same meaning (but not less than potential value).

B.3.4 Does the sentence explicitly elaborate on the meaning of the phrase of interest?

The sentence explicitly elaborates on the meaning of the phrase of interest if it has one of the following relations to the phrase of interest:

1. **Definition** - The sentence establishes rules for what is to be considered an instance of the phrase of interest.
   Examples:
   i. Any mechanical device used for transportation of people or goods is a vehicle.
   ii. FOIA request means a written request for agency records that reasonably describes the agency records sought, made by any person, including a member of the public (U.S. or foreign citizen/entity), partnership, corporation, association, and foreign or domestic governments (excluding Federal agencies).
2. **Explanation** - The sentence explicitly explains certain aspects of the phrase of interest but it does not qualify as a Definition (see the preceding category).
   Examples:
   i. A vehicle usually has wheels, engine and controls.

ii. Likewise, activities are "related" when they are part of a vertical structure such as the manufacturing, warehousing, and retailing of a particular product or products under unified operation or common control for a common business purpose.

3. **Positive Example** The sentence provides a more specific example or an instance of the phrase of interest.

   Examples:

   i. A car is a vehicle.

   ii. The Act defines "record" as any item, collection, or grouping of information about an individual that is maintained by an agency, including, but not limited to, his education, financial transactions, medical history, and criminal or employment history and that contains his name, or the identifying number, symbol, or other identifying particular assigned to the individual, such as a fingerprint or voice print or a photograph.

4. **Negative Example** The sentence provides a specific example or an instance of something that is *not* covered by the phrase of interest.

   Examples:

   i. A stroller is not a vehicle.

   ii. The district court held that Pierce could not satisfy the first prong because duty titles used in lieu of names were not "identifying particulars," and thus the use of duty titles did not make the final response letter and SROI "records" within a "system of records."

5. **Subsumption** The sentence assigns the term or phrase of interest to a more abstract category.

   Examples:

   i. A car is a vehicle.

   ii. Duty titles may indeed be "identifying particulars" as that term is used in the definition of "record" in the Privacy Act.

6. **Contrast** The sentence states that the term or phrase of interest does *not* belong to a certain more abstract category.

   Examples:

i. Not every vehicle is a man-made object.

ii. As explained above, the duty titles do not qualify as identifying particulars.

7. **Feature Assignment** The sentence lists an attribute or a feature of the term or phrase of interest.

   Examples:

   i. Some vehicles are fast.

   ii. The appellee pointed out that duty titles change over time.

8. **Feature Exclusion** The sentence lists an attribute or a feature the term or phrase of interest does *not* possess.

   Examples:

   i. Some vehicles are not large.

   ii. Object code does not derive independent economic value from its secrecy.

B.3.5 Does the sentence provide useful context with respect to the elaboration of the meaning of the phrase of interest?

The sentence provides useful context if one of the relations listed in Subsection B.3.4 could be deduced. By definition any sentence that explicitly elaborates on the meaning of the phrase of interest provides useful context. However, there could also be sentences that do not elaborate explicitly, yet they could be used to derive a similar type of information. Let us consider some implicit versions of the examples listed in Subsection B.3.4:

1. **Definition** - Not applicable.
2. **Explanation** - The vehicle was stripped of its wheels, the engine, and all the controls.
3. **Positive Example** - All the vehicles including the car were parked there.
4. **Negative Example** - Whereas all the vehicles had to be parked in front of the building, the stroller was allowed in.
5. **Subsumption** - We recognized the car among all the vehicles.
6. **Contrast** - Whereas all the vehicles had to be parked in front of the building, the stroller was allowed in.
7. **Feature Assignment** - All the fast vehicles were already gone.

8. **Feature Exclusion** - All the vehicles that were not large could enter the road.

B.3.6 Is the sentence attributed to a person who might have an interest in interpreting the phrase of interest in a certain way?

If the sentence is attributed to a person who has a personal interest in an outcome of legal proceedings, then its value is usually lower. The goal is to determine whether a sentence should be attributed to someone who should be considered as objective (e.g., a judge, a court-appointed expert, etc.) or not (e.g., a party, a witness for one of the parties, etc.). For example, the following sentence would be attributed to a party to the dispute:

> The defendant claimed he did not break the rule since roller skates cannot be considered a vehicle.

APPENDIX C

INTER-ANNOTATOR AGREEMENT REPORT

Term of interest	# sentences	α	# sents	α	α_c	annotator
accommodation trade	69	.40	69	.40	.34	Capybara
					.50	Vulture
audiovisual work	1261	.26	113	.52	.26	Jaguar
					.33	Frog
			434	.27	.43	Raven
					.62	Llama
			714	.06	.20	Seal
					.32	Vulture
aural transfer	139	.24	105	.15	-.52	Seal
					-.11	Herring
			34	.42	-.02	Seal
					.18	Herring
basic allowance for subsistence	79	.47	79	.47	.71	Capybara
					.49	Orca
cybercrime	71	.34	71	.34	.62	Herring
					.48	Llama
dependent on hours worked	27	.56	27	.56	.41	Herring
					.59	Capybara
digital musical recording	43	.32	43	.32	.67	Herring
					.56	Capybara
distributive share of the income	172	.31	172	.31	.41	Llama
					.54	Orca
dischargeable consumer debt	135	-.09	135	-.09	.85	Capybara
					-.04	Vulture
electronic signature	1581	.40	106	.09	.10	Emu
					.43	Herring
			367	.36	.31	Vulture
					.59	Herring
			212	.15	.64	Llama
					.25	Capybara
			319	.39	.45	Raven
					.50	Seal

			577	.52	.59	Orca
					.57	Turkey
essential step	2374	-0.01	210	.24	-.70	Jaguar
					-.70	Vulture
			421	.23	-.63	Jaguar
					-.51	Herring
			428	-.04	-.56	Jaguar
					-.67	Raven
			640	-.29	-.22	Herring
					-.72	Orca
			675	.07	-.62	Vulture
					-.22	Frog
familiar symbol	64	.35	64	.35	.25	Capybara
					.66	Herring
fermented liquor	2133	.42	1112	.61	.75	Llama
					.67	Turkey
			615	.27	.65	Jaguar
					.51	Vulture
			406	.18	.22	Capybara
					.66	Orca
final average compensation	210	.31	210	.31	.50	Raven
					.52	Orca
fully amortize	421	.09	421	.09	.24	Raven
					.32	Orca
gas pipeline facility	66	.11	66	.11	.60	Capybara
					.19	Llama
hazardous liquid	359	.27	359	.27	.21	Raven
					.22	Orca
hybrid instrument	87	.44	87	.44	.03	Capybara
					.40	Vulture
leadership role in an organization	30	.29	30	.29	.90	Herring
					.45	Llama
mechanical recordation	18	.03	18	.03	.40	Herring
					.12	Capybara
navigation equipment	154	.48	154	.48	.64	Llama
					.57	Orca
nonindustrial use	32	-.17	32	-.17	.77	Capybara
					-.02	Herring
nonmonetary benefits	204	.16	204	.16	.31	Capybara
					.26	Orca
preemployment testing	70	.08	70	.08	.57	Capybara
					.27	Llama
preexisting work	823	.56	102	.53	.56	Raven
					.64	Turkey
			310	.50	.56	Capybara
					.73	Herring
			411	.55	.90	Llama
					.62	Seal
residential dwelling	3235	.15	1046	.19	.43	Llama
					.43	Jaguar
			1157	.22	.32	Llama
					.29	Orca

			1032	.05	.39	Capybara
					.25	Herring
security vulnerability	84	.24	84	.24	.62	Capybara
					.15	Orca
semiconductor chip product	25	.50	25	.78	.89	Herring
					.64	Llama
significant property damage	223	-.23	223	-23	.04	Llama
					.20	Orca
small manufacturer	452	.40	108	.58	.28	Capybara
					.00	Vulture
			344	.20	.39	Herring
					.50	Raven
standard coin	179	.35	179	.35	.40	Capybara
					.45	Herring
stored electronically	232	.18	232	.18	.44	Capybara
					.09	Orca
substantial portion of the public	366	-.75	366	-.75	-.67	Llama
					-.02	Orca
switchblade knife	1646	.41	463	.36	.53	Frog
					.41	Turkey
			468	.23	.56	Jaguar
					.58	Orca
			715	.53	.66	Raven
					.51	Capybara
technological measure	616	.03	102	.38	.67	Llama
					.33	Orca
			514	-.08	-.09	Capybara
					.80	Frog
unduly disrupt the operations	48	.03	48	.03	.07	Capybara
					.46	Herring
unreasonably low prices	508	.13	508	.13	.11	Capybara
					.61	Jaguar
useful improvement	3867	.29	1027	.14	.13	Herring
					.42	Capybara
			541	.12	.00	Herring
					.45	Jaguar
			1051	.15	.00	Herring
					.04	Llama
			59	.23	.44	Turkey
					.33	Capybara
			1189	.28	.39	Orca
					.69	Raven
viticultural	221	.31	221	.31	.45	Capybara
					.52	Vulture

APPENDIX D

LEARNING-TO-RANK FEATURES

D.1 INDIVIDUAL UNIT BASED FEATURES

D.1.1 Sentence

1. *Word Count* – The number of words in a sentence.
2. *Exp Word Count* – The number of explanatory words in a sentence.
3. *Exp Word Ratio* – The ratio of explanatory words in a sentence.
4. *Quote Count* – Number of sequences surrounded by quotation marks.
5. *Min Quote Len* – The minimum number of words surrounded by quotation marks.
6. *Max Quote Len* – The maximum number of words surrounded by quotation marks.
7. *Quote Ratio* – The ratio of words surrounded by quotation marks in a sentence.
8. *Introduction* – A binary feature indicating if a sentence comes from the Introduction.
9. *Background* – A binary feature indicating if a sentence comes from the Background.
10. *Analysis* – A binary feature indicating if a sentence comes from the Analysis.
11. *Other* – A binary feature indicating if a sentence comes from the Other.

D.1.2 Query

12. *Qry Word Count* – The number of words in a query.
13. *Root POS NOUN* – A binary feature indicating the POS tag of a query is a NOUN.
14. *Root POS ADJ* – A binary feature indicating the POS tag of a query is an ADJ.

15. *Root POS VERB* – A binary feature indicating the POS tag of a query is a VERB.

D.1.3 Surrounding Sentences

16. *-2s:Word Count* – The number of words in a sentence 2 positions before.
17. *-2s:Exp Word Count* – The number of explanatory words in a sentence 2 positions before.
18. *-2s:Exp Word Ratio* – The ratio of explanatory words in a sentence 2 positions before.
19. *-2s:Quote Count* – Number of sequences surrounded by quotation marks in a sentence 2 positions before.
20. *-2s:Min Quote Len* – The minimum number of words surrounded by quotation marks in a sentence 2 positions before.
21. *-2s:Max Quote Len* – The maximum number of words surrounded by quotation marks in a sentence 2 positions before.
22. *-2s:Quote Ratio* – The ratio of words surrounded by quotation marks in a sentence 2 positions before.
23. *-1s:Word Count* – The number of words in a sentence 1 position before.
24. *-1s:Exp Word Count* – The number of explanatory words in a sentence 1 position before.
25. *-1s:Exp Word Ratio* – The ratio of explanatory words in a sentence 1 position before.
26. *-1s:Quote Count* – Number of sequences surrounded by quotation marks in a sentence 1 position before.
27. *-1s:Min Quote Len* – The minimum number of words surrounded by quotation marks in a sentence 1 position before.
28. *-1s:Max Quote Len* – The maximum number of words surrounded by quotation marks in a sentence 1 position before.
29. *-1s:Quote Ratio* – The ratio of words surrounded by quotation marks in a sentence 1 position before.
30. *+1s:Word Count* – The number of words in a sentence 1 position after.
31. *+1s:Exp Word Count* – The number of explanatory words in a sentence 1 position after.
32. *+1s:Exp Word Ratio* – The ratio of explanatory words in a sentence 1 position after.

33. *+1s:Quote Count* – Number of sequences surrounded by quotation marks in a sentence 1 position after.
34. *+1s:Min Quote Len* – The minimum number of words surrounded by quotation marks in a sentence 1 position after.
35. *+1s:Max Quote Len* – The maximum number of words surrounded by quotation marks in a sentence 1 position after.
36. *+1s:Quote Ratio* – The ratio of words surrounded by quotation marks in a sentence 1 position after.
37. *+2s:Word Count* – The number of words in a sentence 2 positions after.
38. *+2s:Exp Word Count* – The number of explanatory words in a sentence 2 positions after.
39. *+2s:Exp Word Ratio* – The ratio of explanatory words in a sentence 2 positions after.
40. *+2s:Quote Count* – Number of sequences surrounded by quotation marks in a sentence 2 positions after.
41. *+2s:Min Quote Len* – The minimum number of words surrounded by quotation marks in a sentence 2 positions after.
42. *+2s:Max Quote Len* – The maximum number of words surrounded by quotation marks in a sentence 2 positions after.
43. *+2s:Quote Ratio* – The ratio of words surrounded by quotation marks in a sentence 2 positions after.

D.1.4 Paragraph

44. *p:Word Count* – The number of words in a paragraph.
45. *p:Exp Word Count* – The number of explanatory words in a paragraph.
46. *p:Exp Word Ratio* – The ratio of explanatory words in a paragraph.
47. *p:Quote Count* – Number of sequences surrounded by quotation marks.
48. *p:Min Quote Len* – The minimum number of words surrounded by quotation marks.
49. *p:Max Quote Len* – The maximum number of words surrounded by quotation marks.
50. *p:Quote Ratio* – The ratio of words surrounded by quotation marks in a paragraph.

D.1.5 Opinion

52. *o:Word Count* – The number of words in an opinion.

D.1.6 Case

52. *c:Word Count* – The number of words in a case.

D.1.7 Source Provision

53. *prv:Word Count* – The number of words in a source provision.
54. *prv:Exp Word Count* – The number of explanatory words in a source provision.
55. *prv:Exp Word Ratio* – The ratio of explanatory words in a source provision.
56. *prv:Quote Count* – Number of sequences surrounded by quotation marks.
57. *prv:Min Quote Len* – The minimum number of words surrounded by quotation marks.
58. *prv:Max Quote Len* – The maximum number of words surrounded by quotation marks.
59. *prv:Quote Ratio* – The ratio of words surrounded by quotation marks in a source provision.

D.2 QUERY MATCHING BASED FEATURES

D.2.1 Query Matched to Sentence

60. *BM25* – BM25 score between a query and a sentence.
61. *TF* – TF score between a query and a sentence.
62. *Max Qry/Qte Ratio* – Maximum ratio of query terms surrounded by quotation marks.
63. *Dist to Root* – Minimum distance between a query root and a sentence root.
64. *Subtree Size* – Maximum size of a dependency subtree of a query root in a sentence.
65. *Ancestors Size* – Maximum size of a dependency path from a query root to a sentence root.

66. *Exp in Path* – Number of explanatory words in the dependency path between a query root and a sentence root.
67. *Min Dist to Exp* – Minimum length of dependency path between an explanatory word and a query root in a sentence.

D.2.2 Query Matched to Surrounding Sentence

68. *-2s:BM25* – BM25 score between a query and a sentence 2 positions before.
69. *-2s:TF* – TF score between a query and a sentence 2 positions before.
70. *-2s:Max Qry/Qte Ratio* – Maximum ratio of query terms surrounded by quotation marks in a sentence 2 positions before.
71. *-2s:Dist to Root* – Minimum distance between a query root and a root of a sentence 2 positions before.
72. *-2s:Subtree Size* – Maximum size of a dependency subtree of a query root in a sentence 2 positions before.
73. *-2s:Ancestors Size* – Maximum size of a dependency path from a query root to a root of a sentence 2 positions before.
74. *-2s:Exp in Path* – Number of explanatory words in the dependency path between a query root and a root of a sentence 2 positions before.
75. *-2s:Min Dist to Exp* – Minimum length of dependency path between an explanatory word and a query root in a sentence 2 positions before.
76. *-1s:BM25* – BM25 score between a query and a sentence 1 position before.
77. *-1s:TF* – TF score between a query and a sentence 1 position before.
78. *-1s:Max Qry/Qte Ratio* – Maximum ratio of query terms surrounded by quotation marks in a sentence 1 position before.
79. *-1s:Dist to Root* – Minimum distance between a query root and a root of a sentence 1 position before.
80. *-1s:Subtree Size* – Maximum size of a dependency subtree of a query root in a sentence 1 position before.

81. *-1s:Ancestors Size* – Maximum size of a dependency path from a query root to a root of a sentence 1 position before.
82. *-1s:Exp in Path* – Number of explanatory words in the dependency path between a query root and a root of a sentence 1 position before.
83. *-1s:Min Dist to Exp* – Minimum length of dependency path between an explanatory word and a query root in a sentence 1 position before.
84. *+1s:BM25* – BM25 score between a query and a sentence 1 positions after.
85. *+1s:TF* – TF score between a query and a sentence 1 position after.
86. *+1s:Max Qry/Qte Ratio* – Maximum ratio of query terms surrounded by quotation marks in a sentence 1 position after.
87. *+1s:Dist to Root* – Minimum distance between a query root and a root of a sentence 1 position after.
88. *+1s:Subtree Size* – Maximum size of a dependency subtree of a query root in a sentence 1 position after.
89. *+1s:Ancestors Size* – Maximum size of a dependency path from a query root to a root of a sentence 1 position after.
90. *+1s:Exp in Path* – Number of explanatory words in the dependency path between a query root and a root of a sentence 1 position after.
91. *+1s:Min Dist to Exp* – Minimum length of dependency path between an explanatory word and a query root in a sentence 1 position after.
92. *+2s:BM25* – BM25 score between a query and a sentence 2 positions after.
93. *+2s:TF* – TF score between a query and a sentence 2 positions after.
94. *+2s:Max Qry/Qte Ratio* – Maximum ratio of query terms surrounded by quotation marks in a sentence 2 positions after.
95. *+2s:Dist to Root* – Minimum distance between a query root and a root of a sentence 2 positions after.
96. *+2s:Subtree Size* – Maximum size of a dependency subtree of a query root in a sentence 2 positions after.
97. *+2s:Ancestors Size* – Maximum size of a dependency path from a query root to a root of a sentence 2 positions after.

98. *+2s:Exp in Path* – Number of explanatory words in the dependency path between a query root and a root of a sentence 2 positions after.
99. *+2s:Min Dist to Exp* – Minimum length of dependency path between an explanatory word and a query root in a sentence 2 positions after.

D.2.3 Query Matched to Paragraph

100. *p:BM25* – BM25 score between a query and a paragraph in which a retrieved sentence is contained.
101. *p:TF* – TF score between a query and a paragraph in which a retrieved sentence is contained.

D.2.4 Query Matched to Opinion

102. *o:BM25* – BM25 score between a query and an opinion in which a retrieved sentence is contained.
103. *o:TF* – TF score between a query and an opinion in which a retrieved sentence is contained.

D.2.5 Query Matched to Case

104. *c:BM25* – BM25 score between a query and a case in which a retrieved sentence is contained.
105. *c:TF* – TF score between a query and a case in which a retrieved sentence is contained.

D.3 SOURCE PROVISION MATCHING BASED FEATURES

D.3.1 Source Provision Matched to Sentence

106. *New Words* – Number of new words in a retrieved sentence as compared to a source provision.

107. *New Words Ratio* – Ratio of new words in a retrieved sentence as compared to a source provision.

D.3.2 Source Provision Matched to Surrounding Sentences

108. *-2s:New Words* – Number of new words in a sentence 2 positions before as compared to a source provision.
109. *-2s:New Words Ratio* – Ratio of new words in a sentence 2 positions before as compared to a source provision.
110. *-1s:New Words* – Number of new words in a sentence 1 position before as compared to a source provision.
111. *-1s:New Words Ratio* – Ratio of new words in a sentence 1 position before as compared to a source provision.
112. *+1s:New Words* – Number of new words in a sentence 1 position after as compared to a source provision.
113. *+1s:New Words Ratio* – Ratio of new words in a sentence 1 position after as compared to a source provision.
114. *+2s:New Words* – Number of new words in a sentence 2 positions after as compared to a source provision.
115. *+2s:New Words Ratio* – Ratio of new words in a sentence 2 positions after as compared to a source provision.

D.3.3 Source Provision Matched to Paragraph

116. *p:New Words* – Number of new words in a paragraph in which a retrieved sentence is contained as compared to a source provision.
117. *p:New Words Ratio* – Ratio of new words in a paragraph in which a retrieved sentence is contained as compared to a source provision.
118. *p:LDA* – LDA match score between a paragraph in which a retrieved sentence is contained and a source provision.

119. *p:DESMIO* – $\text{DESM}_{I \times O}$ match score between a paragraph in which a retrieved sentence is contained and a source provision.

D.3.4 Source Provision Matched to Opinion

120. *o:LDA* – LDA match score between an opinion in which a retrieved sentence is contained and a source provision.
121. *o:DESMIO* – $\text{DESM}_{I \times O}$ match score between an opinion in which a retrieved sentence is contained and a source provision.

D.3.5 Source Provision Matched to Case

122. *c:fastSIF* – fast^{SIF} match score between a case in which a retrieved sentence is contained and a source provision.
123. *c:LDA* – LDA match score between a case in which a retrieved sentence is contained and a source provision.
124. *c:DESMIO* – $\text{DESM}_{I \times O}$ match score between a case in which a retrieved sentence is contained and a source provision.

D.4 RESULT LIST BASED FEATURES

125. *Number of Cases* – Number of cases in which the retrieved sentences are contained.
126. *Number of Opinions* – Number of opinions in which the retrieved sentences are contained.
127. *Number of Paragraphs* – Number of paragraphs in which the retrieved sentences are contained.
128. *Number of Sentences* – Number of retrieved sentences.
129. *Sentence/Case Ratio* – The ratio between the number of retrieved sentences and the number of cases in which they are contained.

BIBLIOGRAPHY

[1] Words and phrases. https://store.legal.thomsonreuters.com/law-products/Dictionaries-Desk-Reference/Words-and-Phrasesreg/p/100027453. Accessed: 2019-04-15.

[2] James Allan, Courtney Wade, and Alvaro Bolivar. Retrieval and novelty detection at the sentence level. In *Proc. of the 26th international ACM SIGIR conference on Research and development in informaion retrieval*, pages 314–321. ACM, 2003.

[3] Sanjeev Arora, Yingyu Liang, and Tengyu Ma. A simple but tough-to-beat baseline for sentence embeddings. 2016.

[4] Ron Artstein. Inter-annotator agreement. In *Handbook of linguistic annotation*, pages 297–313. Springer, 2017.

[5] Kevin D Ashley. *Modeling legal arguments: Reasoning with cases and hypotheticals*. MIT press, 1991.

[6] Kevin D Ashley. *Artificial intelligence and legal analytics: new tools for law practice in the digital age*. Cambridge University Press, 2017.

[7] B. Bix. *Law, Language, and Legal Determinacy*. Clarendon paperbacks. Clarendon Press, 1995.

[8] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[9] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.

[10] Carlo Bonferroni. Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze*, 8:3–62, 1936.

[11] Abraham Bookstein and Don R Swanson. Probabilistic models for automatic indexing. *Journal of the American Society for Information science*, 25(5):312–316, 1974.

[12] Gerlof Bouma. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL*, pages 31–40, 2009.

[13] Leo Breiman. *Classification and regression trees*. Routledge, 2017.

[14] Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras. Neural legal judgment prediction in english. *arXiv preprint arXiv:1906.02059*, 2019.

[15] G. Chierchia and S. McConnell-Ginet. *Meaning and Grammar: An Introduction to Semantics*. Meaning and Grammar: An Introduction to Semantics. MIT Press, 2000.

[16] Charles Condevaux, Sébastien Harispe, and Stéphane Mussard. Weakly supervised one-shot classification using recurrent neural networks with attention: Application to claim acceptance detection. In *Legal Knowledge and Information Systems: JURIX 2019: The Thirty-second Annual Conference*, volume 322, page 23. IOS Press, 2019.

[17] David Cossock and Tong Zhang. Subset ranking using regression. In *International Conference on Computational Learning Theory*, pages 605–619. Springer, 2006.

[18] Koby Crammer and Yoram Singer. Pranking with ranking. In *Advances in neural information processing systems*, pages 641–647, 2002.

[19] Hang Cui, Min-Yen Kan, Tat-Seng Chua, and Jing Xiao. A comparative study on sentence retrieval for definitional question answering. In *SIGIR Workshop on Information Retrieval for Question Answering (IR4QA)*, pages 383–390, 2004.

[20] Jordan Daci. Legal principles, legal values and legal norms: are they the same or different? *Academicus International Scientific Journal*, 02:109–115, 2010.

[21] Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407, 1990.

[22] Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research*, 7(Jan):1–30, 2006.

[23] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[24] Alen Doko, Maja Stula, and Darko Stipanicev. A recursive tf-isf based sentence retrieval method with local context. *IJMLC*, 3(2):195, 2013.

[25] Harris Drucker. Improving regressors using boosting techniques. In *ICML*, volume 97, pages 107–115, 1997.

[26] Olive Jean Dunn. Multiple comparisons among means. *Journal of the American statistical association*, 56(293):52–64, 1961.

[27] Paul Égré and Nathan Klinedinst. *Vagueness and Language Use*. Palgrave-Macmillan, 2010.

[28] Timothy Endicott. *Vagueness in Law*. Oxford University Press, 2000.

[29] Timothy Endicott. Law and language. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 2016 edition, 2016.

[30] Ronald T Fernández, David E Losada, and Leif A Azzopardi. Extending the language modeling framework for sentence retrieval to include local context. *Information Retrieval*, 14(4):355–389, 2011.

[31] Ronald Teijeira Fernández. *Improving search effectiveness in sentence retrieval and novelty detection*. PhD thesis, Universidade de Santiago de Compostela, 2011.

[32] Alejandro Figueroa and John Atkinson. Contextual language models for ranking answers to natural language definition questions. *Computational Intelligence*, 28(4):528–548, 2012.

[33] John Rupert Firth. A synopsis of linguistic theory 1930–1955. *Studies in Linguistic Analysis*, 1957.

[34] Eibe Frank and Mark Hall. A simple approach to ordinal classification. In *European Conference on Machine Learning*, pages 145–156. Springer, 2001.

[35] Yoav Freund and Robert E Schapire. A desicion-theoretic generalization of on-line learning and an application to boosting. In *European conference on computational learning theory*, pages 23–37. Springer, 1995.

[36] Milton Friedman. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the american statistical association*, 32(200):675–701, 1937.

[37] Milton Friedman. A comparison of alternative tests of significance for the problem of m rankings. *The Annals of Mathematical Statistics*, 11(1):86–92, 1940.

[38] Norbert Fuhr. Some common mistakes in ir evaluation, and how they can be avoided. In *ACM SIGIR Forum*, volume 51, pages 32–41. ACM, 2018.

[39] Debasis Ganguly, Dwaipayan Roy, Mandar Mitra, and Gareth JF Jones. Word embedding based generalized language model for information retrieval. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, pages 795–798. ACM, 2015.

[40] Fredric C Gey. Inferring probability of relevance using the method of logistic regression. In *SIGIR94*, pages 222–231. Springer, 1994.

[41] Harsha Gurulingappa, Luca Toldo, Claudia Schepers, Alexander Bauer, and Gerard Megaro. Semi-supervised information retrieval system for clinical decision support. In *TREC*, 2016.

[42] Jakub Harašta, Tereza Novotná, and Jaromír Šavelka. Citation data of czech apex courts, 2020.

[43] Zellig S. Harris. Distributional structure. *WORD*, 10(2-3):146–162, 1954.

[44] Herbert L. Hart. *The Concept of Law*. Clarendon Press, 2nd edition, 1994.

[45] Stephen P Harter. A probabilistic approach to automatic keyword indexing. part i. on the distribution of specialty words in a technical literature. *Journal of the american society for information science*, 26(4):197–206, 1975.

[46] Yosef Hochberg. A sharper bonferroni procedure for multiple tests of significance. *Biometrika*, 75(4):800–802, 1988.

[47] Matthew Hoffman, Francis R Bach, and David M Blei. Online learning for latent dirichlet allocation. In *advances in neural information processing systems*, pages 856–864, 2010.

[48] Thomas Hofmann. Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 289–296. Morgan Kaufmann Publishers Inc., 1999.

[49] B Holland. On the application of three modified bonferroni procedures to pairwise multiple comparisons in balanced repeated measures designs. *Computational Statistics Quarterly*, 6:219–231, 1991.

[50] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70, 1979.

[51] Gerhard Hommel. A stagewise rejective multiple test procedure based on a modified bonferroni test. *Biometrika*, 75(2):383–386, 1988.

[52] Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*, 2018.

[53] Jerrold Soh Tsin Howe, Lim How Khang, and Ian Ernst Chai. Legal area classification: A comparative study of text classifiers on singapore supreme court judgments. *arXiv preprint arXiv:1904.06470*, 2019.

[54] RL Inman and JM Davenpot. Approximations of the critical region of the friedman statistic. *Communications in Statistics, Theory and Methods A*, 9:571–595, 1980.

[55] Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III. Deep unordered composition rivals syntactic methods for text classification. In *Proceedings*

of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), volume 1, pages 1681–1691, 2015.

[56] Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. Fasttext. zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*, 2016.

[57] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*, 2016.

[58] Matjaz Juršic, Igor Mozetic, Tomaz Erjavec, and Nada Lavrac. Lemmagen: Multilingual lemmatisation with induced ripple-down rules. *Journal of Universal Computer Science*, 16(9):1190–1214, 2010.

[59] Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. A convolutional neural network for modelling sentences. *arXiv preprint arXiv:1404.2188*, 2014.

[60] Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Skip-thought vectors. In *Advances in neural information processing systems*, pages 3294–3302, 2015.

[61] Klaus Krippendorff. Computing krippendorff's alpha-reliability. 2011.

[62] Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. From word embeddings to document distances. In *International Conference on ML*, pages 957–966, 2015.

[63] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019.

[64] Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196, 2014.

[65] Ping Li, Qiang Wu, and Christopher J Burges. Mcrank: Learning to rank using multiple classification and gradient boosting. In *Advances in neural information processing systems*, pages 897–904, 2008.

[66] T.Y. Liu. *Learning to Rank for Information Retrieval*. Springer Berlin Heidelberg, 2011.

[67] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[68] Qiang Lu, Jack G Conrad, Khalid Al-Kofahi, and William Keenan. Legal document clustering with built-in topic segmentation. In *Proceedings of the 20th ACM interna-*

tional conference on Information and knowledge management, pages 383–392. ACM, 2011.

[69] Niklas Luhmann, Klaus A. Ziegert, Fatima Kastner, Richard Nobles, Rosamund Ziegert, and David Schiff. *Law as a Social System*. Oxford University Press, 01 2004.

[70] Sean MacAvaney, Andrew Yates, Arman Cohan, and Nazli Goharian. Cedr: Contextualized embeddings for document ranking. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1101–1104, 2019.

[71] D. N. MacCormick and R. S. Summers. *Interpreting Statutes*. Routledge, Darmouth, 1991.

[72] Gurmeet Singh Manku, Arvind Jain, and Anish Das Sarma. Detecting near-duplicates for web crawling. In *Proceedings of the 16th international conference on World Wide Web*, pages 141–150, 2007.

[73] C.D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[74] Irina Matveeva, Chris Burges, Timo Burkard, Andy Laucius, and Leon Wong. High accuracy retrieval with multiple nested ranker. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 437–444. ACM, 2006.

[75] Samarth Mehrotra and Andrew Yates. Mpii at trec cast 2019: Incoporating query context into a bert re-ranker.

[76] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv:1301.3781*, 2013.

[77] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[78] Tomas Mikolov, Scott Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proc. of the 2013 Conference of the North American Chapter of the ACL: HLT*. ACL, May 2013.

[79] Bhaskar Mitra, Eric Nalisnick, Nick Craswell, and Rich Caruana. A dual embedding space model for document ranking. *arXiv preprint arXiv:1602.01137*, 2016.

[80] Saeedeh Momtazi, Matthew Lease, and Dietrich Klakow. Effective term weighting for sentence retrieval. In *International Conference on Theory and Practice of Digital Libraries*, pages 482–485. Springer, 2010.

[81] Vanessa G Murdock. Aspects of sentence retrieval. Technical report, Massachusetts University Amherst Department of Computer Science, 2006.

[82] Peter Nemenyi. Distribution-free multiple comparisons. In *Biometrics*, volume 18, page 263. International Biometric Soc 1441 I ST, NW, SUITE 700, WASHINGTON, DC 20005-2210, 1962.

[83] Huy Nguyen and Diane Litman. Extracting argument and domain words for identifying argument components in texts. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 22–28, 2015.

[84] Huy V Nguyen and Diane J Litman. Argument mining for improving the automated scoring of persuasive essays. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[85] Quang Nguyen. *Efficient learning with soft label information and multiple annotators*. PhD thesis, University of Pittsburgh, 2014.

[86] Rodrigo Nogueira, Wei Yang, Kyunghyun Cho, and Jimmy Lin. Multi-stage document ranking with bert. *arXiv preprint arXiv:1910.14424*, 2019.

[87] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[88] Matthew E Peters, Waleed Ammar, Chandra Bhagavatula, and Russell Power. Semi-supervised sequence tagging with bidirectional language models. *arXiv preprint arXiv:1705.00108*, 2017.

[89] Jay M Ponte and W Bruce Croft. A language modeling approach to information retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 275–281. ACM, 1998.

[90] The President and Fellows of Harvard University. Caselaw access project. https://case.law/, 2018. Accessed: 2018-12-21.

[91] Tao Qin and Tie-Yan Liu. Introducing letor 4.0 datasets. *arXiv preprint arXiv:1306.2597*, 2013.

[92] Tao Qin, Tie-Yan Liu, Jun Xu, and Hang Li. Letor: A benchmark collection for research on learning to rank for information retrieval. *Information Retrieval*, 13(4):346–374, 2010.

[93] Juliano Rabelo, Mi-Young Kim, and Randy Goebel. Combining similarity and transformer methods for case law entailment. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law*, pages 290–296, 2019.

[94] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. *URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf*, 2018.

[95] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*, 2019.

[96] Jinfeng Rao, Jimmy Lin, and Miles Efron. Reproducible experiments on lexical and temporal feedback for tweet search. In *European Conference on Information Retrieval*, pages 755–767. Springer, 2015.

[97] Jinfeng Rao, Linqing Liu, Yi Tay, Wei Yang, Peng Shi, and Jimmy Lin. Bridging the gap between relevance matching and semantic matching for short text similarity modeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5373–5384, 2019.

[98] Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA.

[99] Stephen Robertson, Hugo Zaragoza, et al. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389, 2009.

[100] Stephen E Robertson and Steve Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *SIGIR94*, pages 232–241. Springer, 1994.

[101] Julien ROSSI and Evangelos KANOULAS. Legal search in case law and statute law. In *Legal Knowledge and Information Systems: JURIX 2019: The Thirty-second Annual Conference*, volume 322, page 83. IOS Press, 2019.

[102] Steven L Salzberg. On comparing classifiers: Pitfalls to avoid and a recommended approach. *Data mining and knowledge discovery*, 1(3):317–328, 1997.

[103] Luis Sanchez, Jiyin He, Jarana Manotumruksa, Dyaa Albakour, Miguel Martinez, and Aldo Lipani. Easing legal news monitoring with learning to rank and bert. Springer, 2020.

[104] Jaromir Savelka and Kevin D Ashley. Extracting case law sentences for argumentation about the meaning of statutory terms. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 50–59, 2016.

[105] Jaromír Savelka and Kevin D Ashley. Segmenting us court decisions into functional and issue specific parts. In *JURIX*, pages 111–120, 2018.

[106] Jaromír Šavelka and Jakub Harašta. Open texture in law, legal certainty and logical analysis of natural language. In *Logic in the Theory and Practice of Lawmaking*, pages 159–171. Springer, 2015.

[107] Jaromir Savelka, Vern R Walker, Matthias Grabmair, and Kevin D Ashley. Sentence boundary detection in adjudicatory decisions in the united states. *Traitement automatique des langues*, 58(2):21–45, 2017.

[108] Jaromir Savelka, Huihui Xu, and Kevin D Ashley. Improving sentence retrieval from case law for statutory interpretation. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law*, pages 113–122, 2019.

[109] D.A. Schmedemann and C.L. Kunz. *Synthesis: Legal Reading, Reasoning, and Writing*. Legal Research and Writing Text Series. Aspen Law & Business, 1999.

[110] Erich Schweighofer. The role of ai & law in legal data science. In *JURIX*, pages 191–192, 2015.

[111] Richard Socher, Eric H Huang, Jeffrey Pennin, Christopher D Manning, and Andrew Y Ng. Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. In *Advances in neural information processing systems*, pages 801–809, 2011.

[112] Roy Sorensen. Vagueness. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 2018 edition, 2018.

[113] Kai Sheng Tai, Richard Socher, and Christopher D Manning. Improved semantic representations from tree-structured long short-term memory networks. *arXiv preprint arXiv:1503.00075*, 2015.

[114] Hamed Valizadegan, Quang Nguyen, and Milos Hauskrecht. Learning classification models from multiple experts. *Journal of biomedical informatics*, 46(6):1125–1135, 2013.

[115] Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 2013.

[116] Vladimir N Vapnik. An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5):988–999, 1999.

[117] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[118] Jesse Vig. A multiscale visualization of attention in the transformer model. *arXiv preprint arXiv:1906.05714*, 2019.

[119] Stephan Walter. Definition extraction from court decisions using computational linguistic technology. *Formal Linguistics and Law*, 212:183, 2009.

[120] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*, 2018.

[121] Wei Wang, Bin Bi, Ming Yan, Chen Wu, Zuyi Bao, Liwei Peng, and Luo Si. Structbert: Incorporating language structures into pre-training for deep language understanding. *arXiv preprint arXiv:1908.04577*, 2019.

[122] Yashen Wang, Heyan Huang, Chong Feng, Qiang Zhou, Jiahui Gu, and Xiong Gao. Cse: Conceptual sentence embeddings based on attention model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 505–515, 2016.

[123] John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. Towards universal paraphrastic sentence embeddings. *arXiv preprint arXiv:1511.08198*, 2015.

[124] Ian H Witten, Eibe Frank, Mark A Hall, and Christopher J Pal. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

[125] Wei Yang, Haotian Zhang, and Jimmy Lin. Simple applications of bert for ad hoc document retrieval. *arXiv preprint arXiv:1903.10972*, 2019.

[126] Zeynep Akkalyoncu Yilmaz, Shengjin Wang, Wei Yang, Haotian Zhang, and Jimmy Lin. Applying bert to document retrieval with birch. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 19–24, 2019.

[127] Zeynep Akkalyoncu Yilmaz, Wei Yang, Haotian Zhang, and Jimmy Lin. Cross-domain modeling of sentence-level evidence for document retrieval. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3481–3487, 2019.

[128] Danchen Zhang and Daqing He. Can word embedding help term mismatch problem?–a result analysis on clinical retrieval tasks. In *International Conference on Information*, pages 402–408. Springer, 2018.

[129] Guangyou Zhou, Tingting He, Jun Zhao, and Po Hu. Learning continuous word embedding with metadata for question retrieval in community question answering. In

Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 250–259, 2015.