Toward Robust and Efficient Interpretations of Idiomatic Expressions in Context

by

Changsheng Liu

Bachelor of Engineering, Huazhong University of Science and Technology, 2008

Master of Science, Peking University, 2012

Submitted to the Graduate Faculty of the

Kenneth P. Dietrich School of Arts and Sciences

in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2019

UNIVERSITY OF PITTSBURGH

KENNETH P. DIETRICH SCHOOL OF ARTS AND SCIENCES

This dissertation was presented

by

Changsheng Liu

It was defended on

March 27 2019

and approved by

Rebecca Hwa, Department of Computer Science, University of Pittsburgh

Adriana Kovashka, Department of Computer Science, University of Pittsburgh

Diane J. Litman, Department of Computer Science, University of Pittsburgh

Yulia Tsvetkov, Language Technologies Institute, Carnegie Mellon University

Dissertation Director: Rebecca Hwa, Department of Computer Science, University of Pittsburgh

Toward Robust and Efficient Interpretations of Idiomatic Expressions in Context

Changsheng Liu, PhD

University of Pittsburgh, 2019

Studies show that a large number of idioms can be interpreted figuratively or literally depending on their contexts. This usage ambiguity has negative impacts on many natural language processing (NLP) applications. In this thesis, we investigate methods of building robust and efficient usage recognizers by modeling interactions between contexts and idioms.

We aim to address three problems. First, how do differences in idioms' linguistic properties affect the performances of automatic usage recognizers? We analyze the interactions between context representations and linguistic properties of idioms and develop ensemble models that predict usages adaptively for different idioms. Second, can an automatic usage recognizer be developed without annotated training examples? We develop a method for estimating the semantic distance between context and components of an idiom and then use that as distant supervision to guide further unsupervised clustering of usages. Third, how can we build one generalized model that reliably predicts the correct usage for a wide range of idioms, despite of variations in their linguistic properties? We recast this as a problem of modeling semantic compatibility between the literal interpretation of an arbitrary idiom and its context. We show that a general model of semantic compatibility can be trained from a large unannotated corpus, and that the resulting model can be applied to an arbitrary idiom without specific parameter tuning.

To demonstrate that our work can benefit downstream NLP applications, we perform a case study on machine translation. It shows that our model can help to improve the translation quality of sentences containing idioms.

Table of Contents

List of Tables

List of Figures

List of Equations

Preface

First and foremost, I want to express my greatest gratitude to my advisor Dr. Rebecca Hwa. I am sure I will fail to precisely describe all the support and assistance I get from you since they are so tremendous. Throughout my PhD study, you gave me valuable freedom to explore the topics I like, while at the same time your expertise in NLP helped me avoid numerous traps and pitfalls along the way I have chosen. We had numerous thought-provoking discussion in our meetings; I got innumerable precious feedback when we worked on papers. Your passion for perfection is infectious. I cannot finish my PhD study without you.

I would also like to thank my dissertation committee: Dr. Adriana Kovashka, Dr. Diane J. Litman and Dr. Yulia Tsvetkov. I have benefited greatly from your refreshing perspective, constructive comments, and generous support. It has been my great privilege to work with you and I appreciate all your time and effort in guiding me through the course of this thesis.

I feel very fortunate to become a member of the NLP group in Pitt. I want to thank all my colleges who have made my experience as a PhD student so pleasant and unforgettable. This especially includes Huichao Xue, Wencan Luo, Lingjia Deng, Mingda Zhang, Homa B.Hashemi, Omid Kashefi, Haoran Zhang and Zahra Rahimi. There are many others in the computer science department who I am fortunate to get to know and become friends with: Rakan Alseghayer, Siqi Liu, Nils Murrugarra Llerena, Michael Cui, Qihang Chen, Zihao Zhao, Yingjie Tang, Jeongmin Lee, Mengsi Lou, Wei Guo, Salim Malakouti, Zuha Agha, Angen Zheng, Keren Ye, Xiaozhong Zhang, Xiaoyu Ge, Yanbing Xue, Duncan Yung, Zhipeng Luo, Kenrick.o.Fernandes, Phuong Pham, Anotoli Shein, Nikos Katsipoulakis and Judicael Briand Djoko,

Finally, I want to extend my sincerest thanks to my parents for your unconditional love. To my girlfriend Yujia Yang and my biggest friend Xiaobing Shi - no matter ups or downs in this journey, you are always there to support and encourage me.

1.0 Introduction

1.1 Motivation

"If natural language had been designed by a logician, idioms would not exist. "

-Johnson-Laird, 1993

Much of human knowledge is contained in and communicated through our languages. Natural Language Processing (NLP) is concerned with developing computational methods to capture this knowledge. A major challenge for computers lies in automating the processing of figurative expressions, such as metaphors and idiomatic expressions. To be able to comprehend the implicit semantics of figurative expressions represents a current frontier in NLP, or more generally, in Artificial Intelligence [Gagliano et al., 2016].

Idiomatic expressions, as a special type of figurative devices, are widely used in different literary genre. A corpus study shows that three out of ten sentences contain idioms [Moon, 1998]. The most significant property of an idiom is that its figurative sense is not simply the combination of the senses of its components (e.g., the idiom "shoot the breeze" means "to chat", which is hard to infer from its component words "shoot" and "breeze"). Idioms often behave idiosyncratically. For example, an idiom may involve the violation of selectional preferences, i.e., a word's tendency to co-occur with words that belong to certain lexical sets (as in "The U.N. is playing with fire"). Meanwhile, idioms exhibit properties of both words and phrases. On the one hand, idioms can be considered as a single fixed semantic unit and their meanings can be comprehended by direct memory retrieval [Cacciari et al., 1993], which are, in a sense, similar to words. On the other hand, idioms are usually multiword expressions, which demonstrate different degree of syntactic flexibility, such as passivizability and tense inflection [Stone and Ann, 2016] (as in "looks like *the beans have been spilled* on one of our portrait artists" or "old motherboard was slowly dying, finally *kicked the bucket* yesterday").

Due to their abundance and idiosyncratic behaviors in natural language, idioms have long been recognized to play a crucial part in NLP. The early literature about the automated processing of idioms was mainly dedicated to *idiom type classification*, which aims to classify whether

an expression is an idiom or not without considering their contexts [Fazly and Stevenson, 2006, Venkatapathy and Joshi, 2005, Katz and Giesbrecht, 2006]. However, idiom type classification is still insufficient for more sophisticated NLP applications. Studies have shown that many idioms can be used *both* figuratively and literally, depending on the context [Fazly et al., 2009]. For example, "break the ice" is used literally in the first instance but figuratively in the second:

(1) When they finally **punched** through the **Arctic** ice cap just shy of the North Pole, it took them five hours to break the ice off their submarine's key hatches so they could reach the fresh air.

(2) US **President** Barack Obama and Cuba's Raul Castro will have a historic face-to-face **encounter** at the **Summit** of the Americas this week, breaking the ice after decades of glacial **relations**.

This ambiguity poses special challenges for various NLP applications. For instance, when we use an advanced information retrieval system to search for some information related to "ice" in physical domain (i.e., the literal sense of "ice"), the system should be able to exclude the second example from our results due to its non-literal interpretation of "ice". In sentiment analysis, idioms have been proven to impose discernible negative impact due to the models' inability to distinguish the literal or figurative senses of idiom [Williams et al., 2015]. In machine translation, previous work [Salton et al., 2014] has shown that a typical statistical machine translation system might achieve only half of the BLEU score [Papineni et al., 2002] on sentences that contain idiomatic expressions than on those that do not.

The inefficiency of NLP applications dealing with this ambiguity makes automatic idiom usage recognition in context, or more generally referred as *idiom token classification*, an indispensable part of NLP research [Fazly et al., 2009]. According to [Jackendoff, 1997], it is estimated that there are 25,000 idioms in the English language alone. Prior efforts on determining the usages of idioms in context fall short when applied to idioms at such a scale. The reasons are twofold:

1. **Large Performance Variance across Idioms**: Different context features have varied predictive power across idioms; while lexical cues are sufficient to distinguish different usages for some idioms (e.g., certain prepositions appearing after "break the ice"), others might need deeper semantic inference. Since these inferences involve processing at different levels of language, it generally requires different representation methods to capture the underlying cues. However, the effectiveness of different representations of context is under-studied in

this task; models proposed in the literature mainly rely on single representation of context and have large performance variances across idioms [Rajani et al., 2014, Birke and Sarkar, 2006, Peng et al., 2014, Sporleder and Li, 2009]. For example, the method of [Fazly et al., 2009] achieves an overall accuracy of 98% on the idiom "take heart", but only gets 35% for the idiom "pull * leg". This performance variance makes current models problematic if we have a large set of idioms to classify. In the context of this thesis, we refer robustness as a model's ability to perform consistently across different idioms.

2. **Intensive Human Labor and Computational Cost**: Idioms vary in form and their contexts of different usages do not follow a set of patterns that can be easily characterized. Hence, a common practice is to train a separate model for each idiom on a large amount of annotated examples [Rajani et al., 2014, Peng et al., 2014]. This is not optimal: (1) annotation needs extensive human effort; (2) a per-idiom model is computationally expensive when we have a large number of idioms. In this thesis, we define efficiency as performing the desired idiom usage recognition task with minimal human supervision and computational resource. To address the efficiency problem, we need either some general knowledge about idiom usages to reduce the need of human supervision, or training the models on generalized features across idioms so that they can be applied to different idioms. As it is hard to find universal patterns from context and idiom **in isolation**, their interactions tend to exhibit some common behaviors across different idioms, e.g., the components of idiom being semantically distant from the context often signals figurative usage. Such types of interactions, while offering a promising opportunity to address the non-efficient problem of current approaches, are less-studied in the literature.

The central goal of this thesis is to address the problems mentioned above and build robust and efficient computational models to recognize an idiom's usage in context.

1.2 Thesis Statement

With appropriate representations of context and idiomatic expression, linguistic-informed computational models which aim to capture the interactions between these representations can help

build robust and efficient idiom usage recognizers. In this thesis, we aim to test the following hypotheses:

H1. Addressing the interaction between context representations and linguistic properties of idioms can help to **train a robust idiom usage recognizer**.

H2. Modeling the interaction between contexts and idioms by calculating their semantic distance and further using it as distant supervision can help to **reduce the need of human supervision**.

H3. Modeling the interaction between contexts and idioms by assessing their semantic compatibility can help to train a generalized model to **reduce computational cost**.

1.3 Thesis Overview

This thesis presents three parts of work with a unifying goal of recognizing an idiom's usage in context. Each part, however, emphasizes a different aspect of the problem. In the first part, we investigate the advantages and limitations of different context representations so as to build more robust idiom usage recognizers by effectively drawing knowledge from these representations. The second and third parts focus on the efficiency problem. To reduce the need of human annotation, we present an unsupervised idiom usage recognizer in the second part. The *semantic similarity* between context and idiom is used as a distant supervision in our proposed models. Continuing on this work, the third part presents a generalized idiom usage recognition model by evaluating the *semantic compatibility* between context and the literal sense of the idiom. The generalized model can reduce the computational cost because there is no need to train the model for each individual idiom. The following is an overview of our work.

1.3.1 Robust Idiom Usage Recognizer

Previous works on idiom usage recognition did not focus on its robustness, so they tend to have large performance variances among different idioms. As noted by [Bengio et al., 2013], the performance of machine learning models is heavily dependent on the choice of data representation. In our

task, we believe that the choice of context representation can significantly impact robustness due to their interactions with linguistic properties of idioms. We summarize the context representations into three main categories: Lexical Representation, Topical Representation and Distributional Semantic Representation. Our studies show that these three representations have different advantages and limitations toward idiomatic expressions. Therefore, how to integrate these representations together and how to incorporate linguistic knowledge of idioms into our model are essential to build a robust idiom usage recognizer. In Chapter 3, we present ensemble models to combine these context representations adaptively for different idioms which can achieve better stability without loss of accuracy.

1.3.2 Unsupervised Idiom Usage Recognizer

Apart from robustness, reducing the need of human supervision is also an important aspect in idiom usage recognition. Most of the success of existing work comes from supervised models, which require human effort to annotate training examples. In this part of the work, we focus on building an idiom usage recognition model without annotated training examples.

Our strategy is to find an alternative form of supervision to automatically replace the supervision signal from human annotation. To achieve this goal, the new form of supervision should be built on features that are abstract enough, such that they are invariant across idioms. For example, lexical features are not optimal since the distribution of context words are specific to each idiom. In Chapter 4, we show how distributional semantic feature comes to the rescue in providing a distant supervision for idiom usage recognizer. We calculate the *semantic similarity* between context and idiom and use this information to guide downstream unsupervised clustering methods, achieving competitive results compared to state-of-the-art supervised models.

1.3.3 Generalized Idiom Usage Recognizer

While reducing human effort in idiom usage recognition is important, reducing the computational cost is also of great significance, considering the number of idioms in language. A particular challenge of automatic idiom usage recognition is that idioms, by their very nature, are idiosyncratic in their usages; therefore, most previous work on idiom usage recognition mainly adopted a

"per idiom" classifier approach, i.e., a classifier needs to be trained separately for each idiomatic expression of interest. In Chapter 5, we propose to build generalized idiom usage recognizers to reduce computational cost. As discussed above, when building an unsupervised model, we model the interaction between idiom and its context by calculating their *semantic similarity* and use it as a type of distant supervision. We push this interaction further by measuring the *semantic compatibility* between context and the literal sense of idiom and use this information to determine the idiom's usage. The concept of *semantic compatibility* is closely related but different with *semantic similarity*: it captures an even more generalized and sophisticated pattern of interaction between a context and an idiom. Our work is based on the observation that most idioms, when taken literally, would be somehow semantically at odds with their context. We have successfully trained a model of semantic compatibility on a large raw corpus and seamlessly apply it to the idiom usage recognition task.

1.4 Contributions

Humans' ability to interpret figurative language, such as inferring the usage of idiom in context, feels so effortless. It can be easy to underestimate how difficult this task is for a computer. The challenge partially lies in the fact that current computational models, as well as the representations of natural language, are not sophisticated enough to capture the complicated semantic relations in language and thus not optimal for high-level semantic tasks such as interpreting figurative language. To understand the nuance and resolve the ambiguity introduced by idiomatic expressions, computers need better semantic representations and more efficient algorithms to make inferences about what they (the idiomatic expressions) are communicating.

From the modeling perspective, this thesis presents models which can recognize the usages of idioms robustly and efficiently.

(1) To build robust idiom usage recognizers [Liu and Hwa, 2017],

- We analyze the advantage and limitation of different context representations quantitatively
- We study two linguistic properties of idioms: semantic analyzability and context diversity. We define two metrics to quantify these properties and explore their interactions with different

representations of context.

- We present an ensemble method based on a variant of averaged perceptron learning method [Collins, 2002] which can effectively integrate different context representations for different idioms.

(2) To build unsupervised idiom usage recognizers [Liu and Hwa, 2018],

- We propose a novel *literal usage metric* based on the semantic similarity between the context and the idiom to estimate the likelihood that the idiom is intended literally.
- We transform the *literal usage metric* into soft labels and present learning algorithms in which the soft label is served as a distant supervision to guild our learning process. We explore two representative probabilistic latent variable models: Latent Dirichlet Allocation (LDA) [Blei et al., 2003] and unsupervised Naive Bayes (NB), in which the usage of idiom is represented as a mixture of linguistically motivated features.

(3) To build generalized idiom usage recognizers [Liu and Hwa, 2019],

- We propose a novel model of *semantic compatibility* by adapting the training of a Continuous Bag-of-Words (CBOW) model for the purpose of idiom usage recognition. The model is trained on a large raw corpus and there is no need to annotate idiom usage examples for training.
- We successfully apply the model on idiom usage recognition and results show that the proposed model achieves competitive results compared to state-of-the-art per-idiom models.

From a practical point of view, our model can alleviate the negative impact caused by idioms in tasks such as machine translation [Salton et al., 2014], sentiment analysis [Williams et al., 2015]. To show the application of our model, we present a case study in which we integrate the usage information of idiom captured by our generalized model into the modern machine translation systems. Results suggest that we can achieve better performance on sentences containing idioms.

The contribution of this thesis is not limited to the automated processing of idioms. First, our linguistically informed ensemble model provides evidence that linguistic is essential to build intelligent and robust models. The experience of our work may serve as an example for bridging the gap between computational models and linguistic theory to other NLP tasks. Second, both the unsupervised learning framework and the semantic compatibility models presented in this thesis

may benefit the NLP community beyond their immediate applications to idiom usage recognition. The idea of soft label as distant supervision can be generalized to other unsupervised learning tasks such as text classification; the concept of semantic compatibility can be applied to detection of other figurative languages such as metaphor and irony.

2.0 Background

In this chapter, we review the literature of research on idioms from a linguistic perspective. We then review the literature of idiom processing from computational perspective. Finally, we describe computational resources (e.g., the shared task and idiom corpora) that are related to this dissertation.

2.1 A Linguistic View of Idioms

Figurative language, such as idiom, metaphor, irony and sarcasm, is ubiquitous in language. Figurative language is generally considered as a creative linguistic device; it is an effective way to convey various meaning such as humor, affection, and express deeply-felt sentiments. As a special type of figurative language, idioms have been studied comprehensively in the linguistic literature. However, there is surprisingly little consensus about the formal definition of idioms. In general, an idiomatic expression can be loosely defined as a combination of words that has a figurative meaning that is hard to infer from the expression's individual components. In this section, we will discuss some linguistic properties of idioms. Since contexts hold clues to resolve the usage ambiguity of idioms, we will have a short discussion about the contexts in which an idiom occurs. Finally, we will briefly review the relationship between idioms and metaphors.

2.1.1 Linguistic Properties of Idiom

Linguists often characterize idioms by certain properties from different perspectives: semantic, syntactic, rhetorical, etc [Nunberg et al., 1994, Cacciari and Levorato, 1998]. We summarize some basic but essential properties of idioms as shown below [Nunberg et al., 1994]:

Conventionality: it refers to the degree to which the figurative meanings of an idiom are not predictable based upon knowledge of its constituents in isolation.

Derivation: the meaning of an idiom might evolve over time. For example, *spill the beans* was

used in horse-racing as early as 1902 and meant ”to cause an upset”. Nowadays, the expression is mainly used to describe the action of revealing a secret.

Inflexibility: the syntactic configurations in which an idiom occurs tend to be relatively fixed. Concretely, an idiomatic expression tends to occur in a small number of canonical form(s). For example, ”break a leg” is a way of wishing good luck before a performance while ”a leg is broken” loses the idiomaticity.

Figuration: idioms often involve metaphor, hyperbole or other types of figuration. For example, the idiom ”I could eat a horse” is an exaggerated way to express that the speaker is extremely hungry. In Section 2.1.3, we discuss the relationship between idioms and metaphors since these two types of figurative language are closely related.

Semantic Analyzability: this measures the extent to which the meanings of the words forming an idiom contribute to its figurative interpretation. Some idioms are completely opaque in terms of semantic, such as ”buy the farm”; a significant amount of idioms are partially transparent, as in ”spill the beans” where ”spill” corresponds to ”divulge” and ”the beans” represents the secret that has been divulged.

What makes idioms interesting and challenging for NLP is that they vary greatly in degrees of these properties. As we have discussed above, idioms have different degree of semantic analyzability. This observation also applies to inflexibility. For example, we have seen that ”break a leg” would lose its idiomatic meaning if it is used in passive voice, whereas some idioms might not, such as ”spill the beans”. Further, it is still an open question to quantify these properties. We find that the measurement of these properties is very subjective and there is no agreed criterion, especially the semantic analyzability. For example, the idiom ”kick the bucket” is generally considered to have low semantic analayzability because the words ”kick”, ”the” and ”bucket” contribute little to its figurative meaning. Nonetheless, [L. Hamblin and Gibbs, 1999] argued that the verb ”kick” conveys a meaning of quickness or suddenness such that ”kick the bucket” means “to die suddenly” rather than “to die slowly.” They suggested that even semantically-opaque idioms are not truly frozen; their figurative meanings are partially shaped by the particular verbs used in these expressions. Due to these reasons, scholars in linguistics have struggled to provide an accurate and predictive model of idiom behaviors. We argue that the automated idiom processing should take the properties of idioms into consideration. We will have more discussion on this point in later

chapters.

2.1.2 The Context of Idiom

The context in which an idiom occurs is essential for determining an idiom's usage. In particular, we find two aspects of context are crucial: Context Word Distribution and Context Diversity.

Context Word Distribution: literal and figurative usages of idiom generally co-occur with different words. More specifically, we find that when an idiom is used literally its contextual words tend to be semantically close to the constituents of the idiom. The intuition is that literal meaning of an idiom is somewhat compositional [Katz and Giesbrecht, 2006]. Literal usages of "get wind", for instance, are more likely to occur with words like "rain", "storm" or "weather" which are related to the constituent "wind". In addition, we find that context word distribution is closely related to semantic analyzability. For idiom with a high degree of semantic analyzability, its figurative meaning is semantically close to its constituent words, thus the overall figurative context would also be close to its literal context.

Context Diversity: this measures how diversified the context of an expression can be. For some idioms, the figurative or literal usage might be closely related to a small range of topics. This is somewhat related to the origin of the idiom. For example, the figurative use of "break the ice" is not very diverse; it is often associated with political topics, so its contexts are likely to contain words such as "country", "nation", "relation" and "war". Other idioms, such as "under the microscope", might be used figuratively with a wider range of topics.

2.1.3 Idioms and Metaphors

Similar to idioms, metaphors are a type of figure of speech which constituent a significant part of human language. A metaphor is formally defined as a conceptual mapping between a source and a target domain [Lakoff and Johnson, 1980]. In other words, it occurs when one concept is regarded as representative or symbolic of another concept from a different domain. For example, consider the metaphor *life is a box of chocolate*, the target (i.e., life) refers to an abstract entity, and the source (i.e., a box of chocolate) refers to a concrete type of food. These two seemly unrelated concepts usually share some hidden similarities so human can build the metaphorical mapping

of the two concepts automatically. Scholars found that this metaphorical mapping preserves the structural characteristics of the source concept [Lakoff, 1990], so people's knowledge of the source concept can help them better understand and conceptualize the target domain.

Metaphors exhibit a great variety, ranging from conventional metaphors, which are commonly used in everyday language, to poetic and creative ones. For conventional metaphors, people use them but pay little attention to which features are mapped from source to target domain, simply because they are widely accepted and become standardized in the language system. On the contrary, it usually requires additional cognitive effort to understand creative metaphors comparing to conventional metaphors [Gibbs Jr, 1992, Gentner and Wolff, 1997, Shutova, 2010b]. [F Bowdle and Gentner, 2005] argued that metaphors undergo an evolutionary path from novel to conventional figurative statements. They referred this path as the "career of metaphor"; the more conventionalized a metaphor becomes, the less thought people pay to its actual mapping.

One widely held belief is that idioms are a type of "dead" metaphors. In other words, they are expressions what were once metaphors but have lost their metaphorical nature over time [Gibbs Jr, 1992]. Early researchers generally assumed that idioms are frozen semantic devices within the speaker's mental lexicons. Their figurative meaning will be retrieved when the literal interpretation is rejected as it is not compatible with the context [A Bobrow and M Bell, 1973]. The dead metaphor view of idiomaticity was questioned by [Gibbs Jr, 1992], who argued that numerous idioms are not "simple, dead metaphors, but actually retain a good deal of their metaphoricity". For example, the figurative meanings of idioms such as *blow your stack*, *flip your lid* and *hit the ceiling* are closely linked to two existing metaphors: MIND IS A CONTAINER and ANGER IS HEATED FLUID IN A CONTAINER. However, the authors only experimented with a small set of idioms, it is not clear whether this argument can still hold when applied to a wider range of idioms. While the relationship between idioms and metaphors is not as straightforward as commonly assumed, they do share the property of being figurative and have overlapping patterns. As such, we review the computational models of metaphor processing in the subsequent section, which provides some context for the studies conducted in this dissertation.

2.2 Computational Background of Idiom Processing

We have now discussed idioms from the linguistic perspective. We have highlighted the properties of idioms and its context. We have also discussed the relationship between idioms and metaphors. The present section provides an overview of computational models of metaphor processing. Since idiom usage recognition can be considered as a type of phrase sense disambiguation problem, which is closely related to word sense disambiguation (WSD), we also present an overview of models of WSD.

2.2.1 Models of Metaphor Processing

Among all different types of figurative language, metaphors have been extensively studied in both NLP and other related fields such as psycholinguistics. We find that the computational models of metaphor processing can be categorized into two groups: metaphor recognition and metaphor interpretation.

2.2.1.1 Metaphor Recognition Metaphor recognition aims to distinguish between literal and metaphorical language in text. The work in this area is pioneer by [Fass, 1991], in which the author presented a system called met* which can recognize metaphor in text using hand-coded patterns such as selectional preference violations. [Mason, 2004] exploited a similar idea to recognize metaphors by finding systematic variations in domain-specific selectional preference. For example, they find that the verb *pour* has a strong selectional preference toward *liquid* in LAB domain but in FINANCE domain it tends to select *money*. Based on this observation they suggest *money* and *liquid* is a metaphorical mapping. As pointed out by Fass, using selectional preference violations as an indicator of metaphor could lead to high false positive; other types of figurative language such as metonymies also frequently involve violations of selectional preferences. Alternatively, [Goatly, 1997] created a set of linguistic cues to recognize metaphors. For example, lexical patterns such as *metaphorically speaking, so to speak* usually signal the presence of metaphorical expression. However, this method suffers from low recall. On the one hand, it is challenging to build a comprehensive set of linguistic cues which are indicators of metaphors. On the other hand,

numerous metaphors occur without explicit linguistic cues.

Apart from selectional preferences violation and linguistic cues, previous work also suggest that the abstractness of context is also an effective indicator of metaphors [Turney et al., 2011, Tsvetkov et al., 2014]. Abstract words refer to things which are hard to perceive directly with our senses. The intuition underlying this line of work is that metaphor can be considered as a method for transferring knowledge from "a familiar, well-understood, or concrete domain to an unfamiliar, less understood, or more abstract domain" [Turney et al., 2011]. Thus, the degree of abstractness in a word's context is correlated with the likelihood that the word is used metaphorically.

2.2.1.2 Metaphor Interpretation Metaphor interpretation aims to explain the intended meaning of metaphorical expressions. Concretely, a large body of work in this area aim to investigate the intuition behind the mapping between the target and source domains. For example, [Kintsch, 2000] proposed a method called Predication Algorithm to find the common features between the target and source domains. Given a predicate P and an argument A (both P and A are represented as vectors in a semantic space using Latent Semantic Analysis [Landauer and Dutnais, 1997]), the method proceeds as follows:

1. Find n closest neighbours of P. Let S denote this set of neighbours.
2. Find k vectors in S that are closest to the argument A and within a threshold t.
3. Calculate the centroid of P, A and the k vectors in step 2.
4. Interpret the centroid by comparing it with a set of suitable landmarks.

Kintsch used "my lawyer is a shark" as an example to illustrate the algorithm. Specifically, the algorithm needs to find the features that are shared by both lawyer and shark. This is challenging because the most obvious features of a shark, e.g., a fish, are not salient to this metaphor. When the centroid is produced in Step 3 above, the paper compared it with six landmarks: three were chosen to be related to lawyer (i.e., lawyer, justice and crime) and three were related to shark (i.e., shark, fish, viciousness). Results suggested the centroid is close to landmark "viciousness" and far from other landmarks, which is an intuitively reasonable interpretation of the metaphor.

[Veale and Hao, 2008] proposed a model called Talking Point, which utilize fluid knowledge representation to interpret metaphor. They build a logic path between the definition of source and target concept by operations such as substitution, insertion and deletion; the logic path is then

The majority of WSD models can be categorized into three main groups: knowledge-based models [Mohammad and Hirst, 2006, Patwardhan and Pedersen, 2006], supervised (semi supervised) models [Mihalcea and Faruque, 2004, Ando, 2006, Zhong and Ng, 2010], and unsupervised models [Agirre et al., 2006, Di Marco and Navigli, 2013].

Knowledge-based models rely on existing lexical resources, such as semantic networks (e.g., WordNet [Fellbaum, 1998]), to identify the most suitable sense. The general idea of this line of work is to measure the relatedness of the senses of the target word to those context words. The most related sense of the target word is selected as the intended sense in the context. For example, [Patwardhan and Pedersen, 2006] used the gloss and structure information in WordNet to build a vector representation for each concept in the WordNet. The sense of a word and the context are represented based on these vectors and their relatedness is measured via cosine similarity.

Since constructing semantic networks is extremely expensive, researchers generally prefer statistical models over knowledge-based models. Supervised statistical models are based on extracting local features from the words surrounding the target, and then training a classifier on annotated examples for each target word. These features include n-grams of nearby words, bag of words, parts-of-speech and syntactic features [Ando, 2006, Zhong and Ng, 2010]. Since sense annotation is time-consuming, semi-supervised methods are often used to alleviate this problem, e.g., a small manually annotated corpus is usually used as seeds for bootstrapping a larger annotated corpus [Mihalcea and Faruque, 2004].

Unsupervised models try to induce word senses directly from the corpus. It is based upon the assumption that similar senses occur in similar contexts, therefore it is possible to cluster word usages according to their context distributions. The clustering algorithms in WSD fall into two categories: vector space model [Pantel and Lin, 2002, Pur and Pedersen, 2004] and graph model [Véronis, 2004, Agirre et al., 2006]. When the clusters (i.e., senses) are induced, the new occurrences of the target word will be compared to these clusters; the most similar cluster will be selected as the intended sense. Note that these methods still need manual intervention to map their induced senses into a sense inventory.

While context is important for WSD in general, many current models ignore the order of words in the context. In the latest development, researcher has used advanced neural networks such as Long Short-Term Memory (LSTM) [Hochreiter and Schmidhuber, 1997] or bidirectional-

LSTM [Graves and Schmidhuber, 2005] to capture the order information in the context for WSD [Kågebäck and Salomonsson, 2016, Pasini and Navigli, 2017]. These methods consider each target word as a separate classification problem and their output layers (i.e., softmax) need to be parameterized for each target word. This practice is similar to "per-idiom" usage recognition models in which we need to train the models for each idiom separately.

It should be clear that the "word-with-spaces" view of idiom is inadequate to account for the complicated behaviors of idioms. Essentially, idioms are different with words in many aspects. As we have discussed in Section 2.1, idioms have their own unique properties such as semantic analyzability and inflexibility. A sophisticated idiom processing model should take these properties into consideration. The above mentioned models of WSD, however, provide important background knowledge for the work in this dissertation.

2.3 Computational Models of Idiom Processing

In NLP, idioms have long been the focus of many research work in the area of figurative language processing. Early research work mainly focus on idiom type classification, i.e., the automatic *identification* of idioms in large corpora. Later research has showed that a potentially idiomatic expression can be used both figuratively and literally, which contradicts the common perception that idioms always have figurative interpretation. We thus see the shift of focus to idiom usage recognition (or idiom token classification) in recent years. In this section, we first briefly discuss the work of idiom type classification. We then review the previous work on idiom usage recognition that our work directly compare with, in which robustness and effectiveness have been largely overlooked.

2.3.1 Idiom Type Classification

Idiom type classification aims to determine whether a phrase could be used as an idiom or not, without considering any specific context. A large body of work in idiom type classification focus on the properties of idioms that differentiate them from the other multiword expressions. These

properties include non-compositionality, lexical fixedness (e.g., ”shoot the breeze” is an idiom but ”shoot the wind” is not) and syntactic fixedness (e.g., ”break a leg” is an idiom but ”a leg is broken” is not). [Tapanainen et al., 1998] proposed a distributed frequency function to determine the non-compositionality in verb-noun collocations. The intuition is that ”if an object only appears with a few verbs, it is highly likely that it has an idiomatic nature”. [Lin, 1999] presented a method to detect non-compositional phrases by checking the mutual information [Church and Hanks, 1990] of phrases obtained by replacing one of the component words with similar words. The main idea is that the mutual information of a non-compositional phrase should differs significantly from the mutual information of phrases obtained by substituting one of the word in the original phrase with a similar word. [Bannard, 2007] presented a method to collect potential idioms by quantifying the syntactic fixedness of phrases. They considered three types of syntactic variations that a phrase can undergo to measure its syntactic fixedness: addition or dropping of a determiner, modification of the noun phrase and passivization of verb phrase. [Fazly and Stevenson, 2006] combined both lexical and syntactic fixedness of verb-noun collocations to gather idioms from large corpora. In terms of lexical fixedness, they measured the deviation between the pointwise mutual information (PMI) [Church et al., 1991] of target expression and the average PMI of its variants (i.e., replacing the noun with similar words). In terms of syntactic fixedness, they use the Kullback Leibler(KL)-divergence between the distributions of the target verb-noun pair and its variants(e.g., passivization and pluralization)

2.3.2 Idiom Usage Recognition

In contrast to idiom type classification, idiom usage recognition (or idiom token classification) aims to determine whether an idiom is meant literally or figuratively in a specific context. A number of idiom usage recognition methods have been proposed: unsupervised [Fazly et al., 2009, Sporleder and Li, 2009, Li and Sporleder, 2009], weakly supervised [Birke and Sarkar, 2006] and supervised [Rajani et al., 2014, Peng et al., 2014].

[Cook et al., 2007] and [Fazly et al., 2009] proposed a method which relies on the concept of *canonical form*. It hypothesizes that in most cases, idioms tend to be somewhat fixed with respect to the syntactic configurations in which they occur, thus idiomatic usages of an expression tend to

occur in a small number of canonical form(s) for that idiom. In contrast, the literal usages of an expression are less syntactically restricted, and are expressed in a greater variety of patterns. This method might be problematic since there are a variety of idiomatic usages of expressions can also occur in non-canonical forms.

[Sporleder and Li, 2009] presented an unsupervised method building on the concept of cohesion graph. They build the undirected graph including all content words in the instances. If removing the idiom improves cohesion, they assume the instance is figurative. Continuing on this work, [Li and Sporleder, 2009] used the unsupervised classifier to label a subset of the test data with high confidence. This subset was then passed on as training data to the supervised classifier, which then labeled the remainder of the dataset.

[Birke and Sarkar, 2006] proposed TroFi (Trope Finder), a weakly supervised method to separate literal and nonliteral usages of verb. It reduced the figurative and literal usage recognition problem to word sense disambiguation by redefining literal and figurative as two different senses of the same word. The core of the algorithm is adapted from word sense disambiguation approach developed by [Karov and Edelman, 1998]. It compares a target expression with two automatically constructed seed sets(one with literal interpretation and one with nonliteral interpretation), assigning the label of the closest set.

When annotated data are available, supervised classifiers are effective. [Rajani et al., 2014] extracted all non-stop-words in the context and used them as ”bag of words” features to train a L2 regularized Logistic Regression (L2LR) classifier [Fan et al., 2008]. [Peng et al., 2014] trained a supervised classifier using the topical features of the context. They also experiment with adding feature of intensity of the emotions in context and find it can provide marginal improvement over the topical feature.

It is worth noting that, apart from the model presented in [Sporleder and Li, 2009], all the models discussed above are not generalized models, which means that they need to be trained for each idiom separately.

2.4 Resource: Shared Task and Corpora

Idioms, or figurative language in general, have become an increasingly prominent part of semantic-oriented applications. In NLP, shared tasks that provide benchmarks for participants to evaluate their systems have greatly boosted the discussion within the community. It is important to emphasize that the quality of the evaluative data is crucial, which is especially true for idioms. Due to their peculiar behavior, it is not as straightforward as other tasks to build appropriate data resources. We think two criteria are important to allow a meaningful evaluation of the success of idiom usage recognizer. First, idioms with highly skewed distribution of figurative and literal usages are not optimal since always choosing the predominant usage can already yield good result. Second, idioms exist in different forms of construction (e.g., verb-noun, noun-noun), so the data should include different types of idioms, or at least include the most representative types which constitute a large amount of idioms. We find three corpora satisfy these requirements: the dataset in SemEval 2013 Task 5B [Korkontzelos et al., 2013], the corpus used in [Fazly et al., 2009] and idiom usage corpus presented in [Sporleder and Li, 2009]. However, the last corpus is not publicly available online, so we mainly use the first two corpora in this dissertation.

2.4.1 SemEval 2013 Task 5B

SemEval (Semantic Evaluation) is an international workshop that conducts evaluations on semantics at different levels. In SemEval 2013 Task 5B, participants were required to make a binary decision whether a target idiom is used figuratively or literally within a given context. For each idiom, several instances extracted from the ukWaC corpus [Baroni et al., 2009] are provided corresponding to its literal and figurative usages. The majority of the instances contains 5 sentences, where the sentence with the target idiom appears in a random position. There are different types of idioms in this task, such as verb-noun combination (V+NN), preposition-noun combination (PP+NN), etc.

In this thesis, we use the following ten idioms from the shared task to evaluate our idiom usage recognition models. These idioms have reasonably large amount of literal and figurative instances, which allows for reliable models to be trained. Note that there are 4 instances labeled as "both"

which could lead to ambiguity are removed and we get 2371 instances in total, among which 1185 instances are literal usages and 1186 instances are figurative.

Table 1: Corpus statistics of SemEval 2013 Task 5b. #Lit denotes the number of literal usages, #Fig denotes the number of figurative usages.

Expression	#Lit	#Fig	All
at the end of the day	102	195	297
bread and butter	148	158	306
break a leg	87	29	116
drop the ball	135	62	197
in the bag	145	156	301
in the fast lane	33	79	112
play ball	157	144	301
rub it in	32	89	121
through the roof	141	170	311
under the microscope	205	104	309

2.4.2 Verb Noun Combination Corpus

A large number of idiomatic expressions are formed by the combination of a verb and a noun (VNC). [Cook et al., 2008] released an idiom usage dateset containing exclusively VNCs. The usage instances are extracted from the British National Corpus (BNC) [Burnard, 2007]. Unlike SemEval corpus, each instance in this corpus contains only 1 sentence. Some idioms from the VNC dataset have very few figurative (or literal) instances, which presents a problem for supervised baselines. To facilitate full comparisons, we select the subset of idioms from the VNC corpus whose number of literal and figurative instances are both higher than 10.

3.0 Context Representations for Robust Idiom Usage Recognizer

3.1 Introduction

Although there are a number of models proposed in the literature which can recognize an idiom's usage in different context, the robustness of these models has received relatively less attention. Reviewing the performance of previous works, we observe that they tend to have large performance variances among different idioms. The objective of this section of the work is to study this problem in depth and investigate the feasibility of building robust idiom usage recognizers.

As noted by [Bengio et al., 2013], the performance of machine learning models is closely related to the choices of data representation. While the local context of an idiom holds clues for discriminating between its literal and figurative usages [Katz and Giesbrecht, 2006], we believe that the choice of context representation can significantly impact robustness of idiom usage recognition. However, the effectiveness of different representations of context is under-studied; we find that models proposed in literature mainly rely on a single representation of context. For example, [Rajani et al., 2014] proposed a supervised model trained on solely on lexical features. As idioms exhibit idiosyncratic behaviors and have varied linguistic properties, relying on a single representation of context is not optimal when applied to a larger set of idioms.

We advocate that in order to fully exploit the information offered by the local context, an idiom usage recognizer ought to leverage knowledge from different types of representation and take the linguistic properties of the idioms into considerations. Among those properties we have reviewed in the background chapter (§ 2.1), we find **context diversity** and **semantic analyzability** significant for usage recognition. Context diversity mainly measures how diversified the context of an expression can be. As we have mentioned previously, if an expression has a low context diversity, a small set of training examples may be sufficient for developing automatic usage recognizer. But for expressions with a high context diversity, however, supervised learning may be unrealistic due to sparsity of training data. Another property semantic analyzability measures the extent to which the meanings of the words forming an idiom contribute to its figurative interpre-

tation [Cacciari and Levorato, 1998]. For idiom with a high degree of semantic analyzability, its figurative meaning is semantically close to its constituent words, thus the overall figurative context would also be close to its literal context. This could make the usage recognition difficult for methods using distributional semantics such as that of [Sporleder and Li, 2009]. Although some previous works do make use of local context, they have not sufficiently taken into account the impact of context diversity and semantic analyzability.

In terms of representations of the context, we find that they can be characterized into three categories: Lexical Representation [Rajani et al., 2014, Birke and Sarkar, 2006], Topical Representation [Peng et al., 2014] and Distributional Semantic Representation [Sporleder and Li, 2009]. Each representation has its own advantages and limitations. Consequently, previous systems tend to perform better for some idioms than others. This thesis hypothesizes that a more flexible and adaptable representation of the context is necessary to account for both context diversity and semantic analyzability. To the best of our knowledge, this work is the first to quantitatively analyze the impact of context diversity and semantic analyzability from a computational perspective. Comparing leading methods against a diverse set of idioms and analyzing the effects of contextual representations, we find that by drawing knowledge from multiple representations and adapting to different idioms, an automatic recognizer can achieve better stability without loss of accuracy.

3.2 Representation of the Usage Context

In this section, we briefly review Lexical Representation, Topical Representation and Distributional Semantic Representation of context. We focus on their limitations and advantages in terms of usage recognition, with an emphasis on their communications with properties of idioms.

3.2.1 Lexical Representation

A straightforward representation is to extract surface words from the context. The assumption is that the contexts of an expression used in the same way should have many words in common. The exact range of the context varies from methods to methods. For example, [Byrne et al., 2013] ex-

tracted only the left and right boundary words of a target phrase to train Naive Bayesian classifiers. On the other hand, [Rajani et al., 2014] extracted all non-stop-words and used them as "bag of words" features to train an L2 regularized Logistic Regression (L2LR) classifier [Fan et al., 2008].

One potential drawback for methods using Lexical Representation is that shared context words are not very strong indicators. Expressions with different usages may nonetheless share some words in common in their contexts; and conversely, even when two contexts do not share any common words, an expression may still have the same usage. Another drawback is that if an idiom has a high degree of context diversity, its contexts would contain too many surface words for them to serve as reliable features.

3.2.2 Topical Representation

Instead of directly setting surface words as the feature space, Topical Representation models a context as a point in an idiomatic expression's topic space. The assumption is that even if an idiom is used in different contexts, if the contexts have similar topics, their usage should be similar. One example of a method in this category is the work of [Li et al., 2010], in which the context is represented as a mixture over latent topics. Another example is the work of [Peng et al., 2014], in which the context is represented as a set of topic words extracted by Latent Dirichlet Allocation (LDA) [Blei et al., 2003].

An advantage of Topical Representation over Lexical Representation is that it could filter out words that are unrelated to the main topics of the context. The discriminative power of words in the context are different; Lexical Representations generally treat all the words equally. Topical Representation extracts the most critical words for the relevant topics. It can be seen as a refined version of Lexical Representation. For example, Topical Representation would extract the most informative words such as *Freedom*, *Democracy* and *President* in the following sentence to help determine the usage of *break the ice*. These words are generally related to political topics, indicating *break the ice* is more likely to be used figuratively.

> (1) **President** Obama, who started his approach toward the radical Islamists ruling Iran by extending a hand, turning his back on the Iranian people with their aspirations for **freedom** and **democracy**, hoped that he could be the first U.S. **president** to break the ice with the Jihadists in Tehran.

A possible drawback of Topical Representation is that it might overlook some syntactic informa-

tion which could be used in the usage recognition for some idioms. For example, a figurative usage of *break the ice* may be indicated by the occurrence of the prepositions *over* or *between* after it [Li and Sporleder, 2010]. These words are generally ignored by methods using Topical Representations, whereas methods using Lexical Representation may include them. Also, similar to Lexical Representation, the context diversity will also influence the effectiveness of Topical Representation.

3.2.3 Distributional Semantic Representation

Methods using the previous two representations essentially rely on the calculation of common words between contexts, which is problematic for idioms with a high degree of context diversity. Distributional Semantic Representation can overcome this problem by using external resource or knowledge base to calculate words similarity. For instance, the following sentence has no word overlap with example (1). However, the word *monarch* is semantically close to the word *president* in example (1), which suggests they might have the same usage.

(2) Edwards usually manages to break the ice with the taciturn **monarch**.

One method in the literature that used Distributional Semantic Representation is the work of [Sporleder and Li, 2009]. They used distributional semantic similarity to calculate the lexical cohesion [Halliday and Hasan, 2014] between constituent words of an idiom and its contextual words. The hypothesis of this method is that if the constituents of a potentially idiomatic expression do not 'fit' in any lexical chains, it is highly likely that the expression is used figuratively.

Despite its advantage, Distributional Semantic Representation still has its limitations. First, for some idioms, it is more effective to just use the surrounding words to detect its usage, such as the preposition *over* or *between* after *break the ice*. Second, since the approach assumes that the overall literal context and figurative context is semantically distant, it is poor at handling idioms with a high degree of semantic analyzability.

3.3 Our Model

We treat literal and figurative usage recognition as a special word sense disambiguation problem in the same spirit as [Birke and Sarkar, 2006]. Specifically, we use similarity-based models because they have been shown to be effective in the general problem of word sense disambiguation [Abdalgader and Skabar, 2012, Karov and Edelman, 1998]. In this section, we describe two variants of our model for integrating different contextual representations within our similarity-based framework.

Representation fusion strategies To fuse different context representations, one straightforward strategy is to concatenate all the features using the three representations and build a single similarity based classifier that applies to the concatenated feature (*early fusion*) [Bruni et al., 2014]. Another option is a per-representation strategy; different classifiers are trained independently on the three representations, and afterwards, the results are combined to generate a final output (*late fusion*). We have experimented with both strategies.

3.3.1 The Late Fusion Model

In this model, three classifiers are developed based on Lexical similarity, Topical similarity and Distributional semantic similarity; and a variant of averaged perceptron learning is applied to learn the weights for each classifier according to its discriminative power over different idioms.

Lexical similarity: Given two contexts T_i and T_j of a target expression, we use cosine similarity to calculate their similarity as shown in the Equation 4.3, where T^i_{bow} and T^j_{bow} denote the bag of word vector of the two contexts. We remove all the stop words in the context except the preceding and following words of the target expression, which tend to be useful for some idioms [Byrne et al., 2013].

$$Sim_1(T_i, T_j) = \frac{T^i_{bow} \cdot T^j_{bow}}{|T^i_{bow}| \cdot |T^j_{bow}|} \tag{3.1}$$

Topical similarity: For an idiom, we first run LDA to all the instances and get a set of *m* topics. For each instance, we represent the context using its probabilities over these topic set. Given two contexts T_i and T_j, we use T^i_{topic} and T^j_{topic} to denote their Topical Representations. Their topic similarity is calculated also using cosine similarity.

$$Topics = \{t_1, t_2, ..., t_m\} \tag{3.2}$$

$$T_{topic} = \{P(t_1), P(t_2), ..., P(t_m)\} \tag{3.3}$$

$$Sim_2(T_i, T_j) = \frac{T^i_{topic} \cdot T^j_{topic}}{|T^i_{topic}| \cdot |T^j_{topic}|} \tag{3.4}$$

Distributional semantic similarity: Given two contexts T_i and T_j, we calculate their semantic similarity $Sim_3(T_i, T_j)$ using doc2vec [Le and Mikolov, 2014]. In detail, we use gensim toolkit [Řehůřek and Sojka, 2010] and train our model on Wikipedia articles[1]. We empirically set the dimensionality of vector to 200.

$$Sim_3(T_i, T_j) = doc2vec\_sim(T_i, T_j) \tag{3.5}$$

We distinguish the usage of the target expression by calculating its average similarity (using one of the similarity metrics) to both the literal and figurative example set and assign the label of the set which has higher similarity. Since we have three types of similarity metrics, we now have three "voters". We use v_i to denote the voting vector with each entry representing the voting results for the ith instance of a idiom.

Because idioms vary in properties that may impact each representation differently, we propose to learn the weight for each voter by applying a variant of averaged perceptron learning method [Collins, 2002]. In addition, we augment the weight learning algorithm by incorporating a novel confidence measure [Schapire and Singer, 1999]. In our case, the confidence is related to the similarity difference. Let Sim_f be the similarity between the context of the target expression and figurative example set, Sim_l be the similarity between the context of the target expression and literal example set (using any of the three similarity metrics). The ratio between the two similarities is a reasonable confidence measure at first glance. The intuition is that the bigger the difference between the two similarities Sim_f and Sim_l, the more confident the voter is. However, both our empirical evidence and observation from [Schapire and Singer, 1999] suggest such confidence measure could lead to large and overly confident predictions and ultimately increases the possibility of overfitting. To overcome such issue, we use a smoothed ratio between the two similarities as the confidence value shown in Equation 3.6. Similar to voting vector v_i, we construct the confi-

[1] https://dumps.wikimedia.org/enwiki/latest/enwiki-latest-pages-articles.xml.bz2

$$c = 1 + \ln \frac{max(Sim_f, Sim_l)}{min(Sim_f, Sim_l)} \tag{3.6}$$

dence vector c_i for the ith instance; the confidence rated voting vector x_i is the point-wise product of v_i and c_i. Then we apply the voting weight learning algorithm to get the weight w for each voter and classify the target expression usage using Equation 3.7.

$$y^* = \text{sign}(wx_i) \tag{3.7}$$

3.3.2 The Early Fusion Model

In this case, we perform L-2 normalization and simply concatenate the vectors of the three representations and then apply cosine similarity metric. The classification process is identical to the single classifier in late fusion strategy.

3.4 Experiment

To verify our hypothesis that robust idiom usage recognition depends on addressing the interactions between properties of idioms (i.e., context diversity and semantic analyzability) and contextual representations, we conduct a comparative study across four representative state-of-the-art methods: two for Lexical Representation [Rajani et al., 2014, Birke and Sarkar, 2006][2]; one for Topical Representation [Peng et al., 2014]; and one for Distributional Semantic Representation [Sporleder and Li, 2009]. We then compare our proposed methods against these four. The experiments address the following questions:

- To what extent can usage recognizers reliably predict figurative versus literal usages for a wide variety of idioms?
- For some of the comparative methods, contextual information is only a portion of many other features, what is the relative contribution from contextual information compared to other features?

[2] We include Rajani et al.'s method because it achieves the best performance on the SemEval 2013 task 5B corpus.

- Does our proposed model of adapting multiple contextual representations succeed in capturing the interactions between representational choices and context diversity and semantic analyzability?

Evaluative Data We use the SemEval 2013 Task 5B corpus described in Chapter 2. We do not use the VNC corpus in this study because the number of instances for each idiom in VNC corpus is not enough for us to quantitatively measure its linguistic properties. On average, each idiom in SemEval corpus has more 200 instances, while the idioms in VNC corpus only have about 60 instances.

Evaluation Metric We rely on the standard F1 score for the recognition of the figurative usage. The overall accuracy of both figurative and literal usage is not ideal for analysis because it can be misleading for idioms with unbalanced usage distribution.

3.4.1 Implementation

We reimplemented the four methods with two minor changes. First, Sporleder and Li used Normalized Google Distance (NGD) to measure the semantic relatedness between two words [Cilibrasi and Vitanyi, 2007], but the API of NGD has a restriction on the number of queries it can make; therefore, we use word embeddings for calculating the distributional semantic similarity [Mikolov et al., 2013b]. Second, we did not encode Birke and Sarkar's SuperTags feature because they reported that the overall gain was only 0.5%. We do not expect these two changes to have significant impact on the findings.

We run ten fold cross validation for the supervised methods (Rajani et al., Peng et al. and our full models). In each round of the cross validation, we randomly select half of the training sample as the example set; the remaining half of the training sample is used to learn the weight for the three representations.

3.4.2 Results and Observations

Table 3 reports the performances of the four comparative state-of-the-art methods. As expected, the supervised classifier by Rajani et al.'s achieves the best performance while the unsupervised method by Sporleder and Li has the lowest scores for most idioms.

Comparing across different idioms for each method, we observe large performance variances. For Rajani et al., the F_{fig} is as low as 0.54 for *break a leg* and as high as 0.83 for *through the roof*. Similarly, Peng et al., the lowest F_{fig} is 0.46 for *under the microscope* and the highest is 0.75 for *at the end of the day*.

Table 4 shows the performances of the two supervised methods limited to just the contextual features. Compared to their full model counterparts in Table 3, we see that the contribution from the additional features is limited, and its impact varies from idiom to idiom. For some, the additional features might have negative effect on the performance (cf. *in the bag*). These results suggest that contextual features are essential to the idiom usage recognition task.

Table 5 reports the performances of our proposed models (both early fusion and late fusion), each of the three component representations in the late fusion model, and the best of the comparative methods for each idiom. The performance of our full late fusion model is competitive; most of our F_{fig} are higher than the best results from the other methods. The late fusion model is more stable than the other methods, with a narrow range of F_{fig} scores, from 0.68 (*under the microscope*) to 0.85 (*at the end of the day*).

3.4.3 Discussion: Performance Variance

We have hypothesized that the variance in performance is partially due to context diversity. In general, methods using surface representation (Rajani and Birke) expect a large training set or seed set with a good distribution which could include sufficient decisive contextual words for a given target expression. This also applies to methods using topic representation since text is modelled as a mixture over latent topics, which are also represented by a distribution over **word**. For some idioms, the figurative or literal usages might be closely related to a small range of topics. Take the idiom *break the ice* as an example, it has a figurative meaning:

to relax a tense or unduly formal atmosphere or social situation.

This is most frequently used in political topics. So the figurative cases are often found in a context containing words such as **country**, **nation**, **relation**, and **war**. To train a recognition model for this type of idioms, even a small amount of training examples could be sufficient to capture a fairly complete semantic features. However, we note that it is infeasible to annotate enough number of

Table 3: Result of different methods. F_{fig} denotes F1 score of figurative usage recognition and A denotes the overall accuracy. For each idiom, the boldfaced number shows the best performance among the four methods while underlined shows the worst.

	Rajani et al.		Peng et al.		Sporleder and Li.		Birke and Sarkar	
Idiom	F_{fig}	A	F_{fig}	A	F_{fig}	A	F_{fig}	A
at the end of the day	**0.81**	**0.73**	0.75	0.63	0.72	<u>0.59</u>	<u>0.69</u>	0.63
bread and butter	**0.81**	**0.8**	0.75	0.70	<u>0.66</u>	<u>0.58</u>	0.67	0.70
break a leg	0.54	**0.8**	<u>0.49</u>	<u>0.63</u>	**0.67**	0.7	0.61	0.65
drop the ball	**0.61**	**0.79**	0.58	0.67	<u>0.45</u>	<u>0.32</u>	0.52	0.76
in the bag	**0.72**	**0.71**	0.68	0.66	0.65	<u>0.50</u>	<u>0.64</u>	0.71
in the fast lane	**0.78**	0.67	0.72	**0.69**	<u>0.52</u>	<u>0.61</u>	0.68	0.65
play ball	**0.75**	0.72	0.68	0.67	<u>0.51</u>	<u>0.40</u>	0.73	**0.75**
rub it in	**0.67**	**0.69**	0.5	0.47	0.55	<u>0.46</u>	<u>0.44</u>	0.49
through the roof	**0.83**	**0.81**	0.68	0.69	<u>0.61</u>	<u>0.51</u>	0.69	0.74
under the microscope	**0.55**	0.74	0.46	0.64	<u>0.42</u>	<u>0.41</u>	**0.55**	**0.79**

Table 4: Result of two supervised methods using only contextual features. F_{fig} denotes F1 score of figurative usage recognition and A denotes the overall accuracy.

	Rajani et al.		Peng et al.	
Idiom	F_{fig}	A	F_{fig}	A
at the end of the day	0.8	0.71	0.73	0.61
bread and butter	0.85	0.84	0.74	0.69
break a leg	0.57	0.77	0.46	0.60
drop the ball	0.59	0.77	0.59	0.68
in the bag	0.75	0.75	0.66	0.62
in the fast lane	0.78	0.68	0.68	0.64
play ball	0.84	0.82	0.64	0.61
rub it in	0.66	0.67	0.51	0.49
through the roof	0.78	0.77	0.67	0.62
under the microscope	0.5	0.74	0.51	0.66

examples for some idioms since they can be used in a wide variety of topics, among which their semantic context could be significantly different with each other.

To measure the diversity of contextual words for a target idiom is essentially similar to measuring the diversity of topics in which the idiom can be used. We can manually annotate each example using a predefined topic set. Nevertheless, it's difficult to define a topic set with appropriate granularity. A small set of high level topics is too general to distinguish different examples and thus cannot fully assess the diversity of topics. On the other hand, a large set of specific topics can lead to an inflated diversity measurement. It also might result in low inner annotation agreement since an example can be labelled with different topics if the topic set is too detailed. In addition, it's labor intensive to annotate all the examples. LDA is a potential method to automatically generate the set of topics based on probability which maybe more desirable. So alternatively, we run LDA method to the examples for a given idiom by varying the number of topics. For each topic number, a log-likelihood value is calculated, indicating how well the generated topic model fits the example set. We select the number of topics with the highest log-likelihood value to approximate the measurement of diversity of topics for the idiom (see Formula 3.8, D denotes the example set, M_n denotes the generated model with n as the topic number).

$$\underset{n}{\operatorname{argmax}}\ logP(D|M_n) \tag{3.8}$$

We randomly select 32 literal instances and 29 figurative instances (the minimum number of instances among all the target idioms) for each idiom from the corpus and run the process mentioned above. The results are shown in Table 6.

We observe that *under the microscope* has the highest topic number, suggesting that it has a high context diversity; it is an idiom that is difficult for all four methods. In contrast, the optimal topic numbers for *bread and butter* is the lowest, suggesting that it has a low context diversity; accordingly, methods using Lexical Representation and Topical Representation performed well on it. We also calculate the Pearson correlation between F_{fig} and the total topic number.[3] The r value is -0.86 for Rajani et al., which suggests strong negative correlation; while the r values for Peng et al. and Birke and Sarkar are -0.72 and -0.62 respectively, suggesting a more moderate

[3] For methods from Rajani et al. and Peng et al, we use the F_{fig} from Table 4 (the implementation without additional features).

Table 5: The comparison between our method and competing methods. The "Best other" column shows the best result from the other methods. $*$ indicates the difference between the "Late fusion" and "Best other" is statistically significant, χ^2 text, p = 0.05. The boldfaced number shows the best performance.

	Best other		Lexical		Topical		Distributional		Early fusion		Late fusion	
Idiom	F_{fig}	A	F_{fig}	A	F_{fig}	A	F_{fig}	A	F_{fig}	A	F_{fig}	A
at the end of the day	0.81	0.73	0.82	0.75	0.81	0.74	0.72	0.69	0.79	0.73	**0.85***	**0.81***
bread and butter	0.81	0.8	0.83	0.79	**0.84**	0.80	0.57	0.61	0.82	0.71	**0.84**	**0.83**
break a leg	0.67	0.7	0.58	0.7	0.56	0.63	0.69	0.71	0.66	0.7	**0.73***	**0.71**
drop the ball	0.61	0.79	0.65	0.81	0.59	0.77	0.51	0.69	0.67	0.82	**0.72***	**0.85***
in the bag	0.72	0.71	0.67	0.66	0.67	0.69	0.74	0.71	0.73	0.65	**0.75***	**0.74**
in the fast lane	**0.78**	0.67	0.68	0.69	0.70	0.73	0.59	0.65	0.54	0.69	0.72*	**0.74***
play ball	0.75	0.72	0.76	0.77	0.71	0.76	0.61	0.71	0.78	0.74	**0.82***	**0.81***
rub it in	0.67	0.69	0.65	0.68	0.73	0.71	0.62	0.71	0.7	0.71	**0.78***	**0.76***
through the roof	**0.83**	0.81	0.81	0.8	0.71	0.69	0.65	0.72	0.81	0.66	0.81	**0.85**
under the microscope	0.55	**0.79**	0.64	0.73	0.47	0.66	0.52	0.69	0.58	0.75	**0.68***	0.75

negative correlation. Although the r value for Sporleder and Li is -0.72, which also suggests a moderately negative correlation, its trend is less reliable. For example, *through the roof* has the lowest topic number (12), but the F_{fig} score (0.61) is well below the best result (0.72); *break a leg* has a relatively high topic number (18), but the F_{fig} score (0.67) is better than the other three methods. These observations suggest that context diversity does influence performances, especially for methods using Lexical or Topical Representation.

Performance variance may also be due to semantic analyzability, especially for methods using Distributional Semantic Representation. We quantify semantic analyzability in the following way. For an idiom, we prepare two sets of instances; one consists of literal instances and the other consists of figurative instances. Then we approximate the semantic analyzability of the idiom by measuring the averaged semantic similarity between the two sets. We use L and F to represent the literal and figurative set respectively. The averaged similarity of F and L is calculated using the following Formula:

$$S_{set}(F, L) = \frac{1}{|F|} \sum_{\forall T_f \in F} \max_{\forall T_l \in L} doc2vec_{sim}(T_f, T_l) \tag{3.9}$$

Table 7 shows our semantic analyzability measure on the 10 idioms. The idiom with the highest similarity score is *drop the ball*, indicating that literal and figurative usages are hard to separate. This corresponds to the poor performance of Sporleder and Li's method on it. In contrast, *break a leg* has the lowest similarity score, which corresponds to the high F_{fig} using Sporleder and Li's method. We also calculate the Pearson correlation coefficient between the F_{fig} and $S_{set}(F, L)$; the r value is -0.77 for Sporleder and Li's method, which suggests moderate negative correlation between the two variables; the r values for the other three methods are -0.03, 0.17, 0.06, respectively. These findings lend credence to our argument that semantic analyzability influences the effectiveness of Distributional Semantic Representation.

Table 6: Optimal topic numbers for different idiom instances. T_{Fig} means the topic number of figurative set, T_{Lit} means the topic number of literal set.

Idiom	T_{Fig}	T_{Lit}	Total
at the end of the day	9	4	13
bread and butter	7	5	12
break a leg	12	6	18
drop the ball	13	8	21
in the bag	11	6	17
in the fast lane	9	7	16
play ball	9	7	16
rub it in	12	5	17
through the roof	8	4	12
under the microscope	16	7	**23**

3.4.4 Discussion: Combining Different Representations

Throughout this chapter, we have argued for the importance of combining different representations of the context. As shown in Table 5, the stability of the late fusion model did improve. But do the results of the individual components corroborate our arguments about the interactions between linguistic properties and specific representations?

Consider *break a leg*, which has a higher context diversity (18 topics) but lower semantic analyzability (0.27 similarity score). Our model's Lexical Representation and Topical Representation components are not as effective as the Distributional Semantic Representation component; they

Table 7: A measure of Semantic Analyzability

Idiom	Similarity
at the end of the day	0.28
bread and butter	0.32
break a leg	0.27
drop the ball	**0.37**
in the bag	0.29
in the fast lane	0.35
play ball	0.34
rub it in	0.28
through the roof	0.32
under the microscope	0.34

have an F_{fig} score of 0.58, 0.56, and 0.69 respectively. Similarly, for an idioms with a higher semantic analyzability but a lower context diversity like *bread and butter*, our model's Distributional Semantic Representation component performed worse individually than the Lexical Representation and Topical Representation components.

In both cases, our method has effectively adapted to the particulars of the idioms and increased the contributions from the well performing components. For *break a leg*, the weights of the components are $[0.23, 0.19, 0.58]$, favoring the Distributional Representation to obtain an F_{fig} of 0.73. For *bread and butter*, the weights are appropriately shifted to the Lexical Representation and Topical Representation components ($[0.4, 0.43, 0.17]$) for an overall F_{fig} of 0.84.

3.4.5 Limitations

We have discussed the interaction between linguistic properties of idioms and context representations. However, some limitations should be noted. First, the number of idioms used in the experiment is fairly small due to the expensive cost of data collection and usage annotation. Including more idioms would make the conclusion more reliable. A second limitation concerns the measurement of the linguistic properties, especially semantic analyzability. Although this property has been discussed intensively in the linguistic literature, there is no standard way to represent it quantitatively; we use the averaged semantic similarity between the literal and figurative instances

as an indirect measurement. As we can see from Table 7, the range of semantic analyzability is somewhat narrow, which might not accurately reflect the distinction among idioms. An alternative solution is to manually rate the semantic analyzability of idioms. However, as this property is highly subjective, it is therefore prone to rater bias.

3.5 Chapter Summary

To build a robust idiom usage recognizer, we have argued for the importance of two linguistic properties in idioms (context diversity and semantic analyzability) and analyzed their impact on context representations. Experimental results show that leading methods with fixed representations do not perform equally well on different types of idioms. We have proposed a supervised ensemble approach to adaptively combine multiple contextual semantic representations for different idioms. Evaluated on a diverse set of idioms, we find that our method can achieve better stability without loss of accuracy.

4.0 Heuristically Informed Unsupervised Idiom Usage Recognition

4.1 Introduction

We have presented a supervised method for building a robust idiom usage recognizer in the previous chapter. The model requires appropriately annotated examples, which is time-consuming. To process idioms on a large scale, effectiveness is as important as robustness. In this chapter, we focus on reducing the need of human supervision in idiom usage recognition.

Some previous unsupervised models tried to exploit linguistic differences in usages. For example, [Fazly et al., 2009] observed that an idiom appearing in its canonical form is usually used figuratively; [Sporleder and Li, 2009] relied on the break in lexical coherence between the idioms and the context to signal a figurative usage. These heuristics, however, are not always applicable because the distinctions they depend upon may not be present or obvious. To improve generalization across different idioms and usage contexts, we need a more reliable heuristic, and appropriately incorporate it into an unsupervised learning framework.

We propose a novel heuristic that differentiates an idiom's usages based on distributional semantics [Harris, 1954, Turney and Pantel, 2010]. Our key insight is that when an idiom is used literally, its relationship with its context is more predictable than when it is used figuratively. This is because the literal meaning of an idiom is compositional [Katz and Giesbrecht, 2006], and the constituent words that make up the idiom are also meant literally. For example, in the following sentence,

***Spill the beans**, flip the fruit, bust open a box of hot pockets. Make a general mess of the kitchen.*[1]

spill is meant literally and can take on objects other than *beans*; moreover, one of the context words, *mess*, can often be seen to co-occur with *spill* in other text, even without *beans*. Our strategy is to represent an idiom's literal usage in terms of the word embeddings of the idiom's constituent words and other words they frequently co-occur with. Then, for any instance in which the idiom's usage is not known, we only need to determine the semantic similarity between that instance and

[1] https://twitter.com/DukeRaccoon/status/477530732173471744

the idiom's literal representation. We expect a high similarity score generally indicates a high probability of literal usage. The raw scores may be difficult to interpret since different idioms can have wildly varying score ranges. We propose a *literal usage metric* which transforms the raw scores into a probabilistic interpretation – the likelihood that an instance would be labeled "literal". Having a metric with a probabilistic interpretation also affords us a greater flexibility in terms of using it to inform downstream learning processes.

While the literal usage metric captures the distributional semantic information of the context, we find that some other linguistic cues are also significant for usage detection (such as whether the subject of the sentence is a person); therefore, we allow our model to further refine through unsupervised methods. Specifically, we treat the usage (*figurative* or *literal*) as a hidden variable in probabilistic latent variable models, and we define a set of features that are linguistically relevant for idiom usage detection as observables. We integrate our literal usage metric with the latent variable models by treating the metric outputs as *soft labels* to guide the latent variable models toward grouping by usages.

We hypothesize that unsupervised learning in a more linguistically motivated feature space, informed by soft labels from a semantically driven metric, will produce more robust classifiers. We conduct experiments comparing our approach against other supervised and unsupervised baselines. Results suggest that our approach achieves performances that are competitive to supervised models.

4.2 Our Approach

Given a target idiomatic expression and a collection of instances in which the idiom occurs, our proposed system determines whether the idiom in each instance is meant figuratively or literally, without using idiom specific resources such as a dictionary or an annotated corpus.

An overview of our approach is illustrated in Figure 2. We first build a **Literal Usage Representation** for each idiom by leveraging the distributional semantics of its constituents (§ 4.2.1). Given an instance of idiom, we can determine its usage by the semantic similarity between the context of the instance and the **Literal Usage Representation**. We define a **Literal Usage Metric**

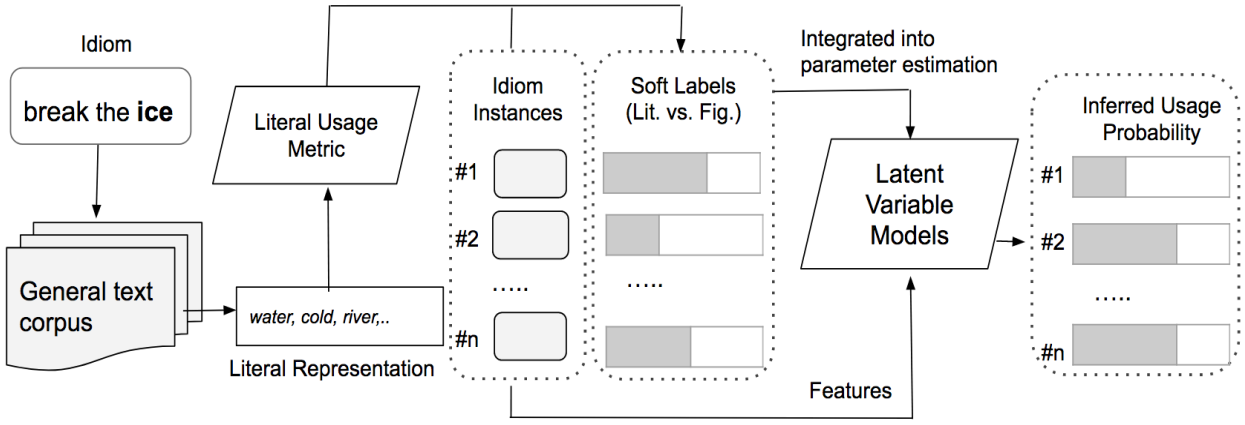


Figure 2: An overview of our unsupervised idiom usage recognition model.

to transform the semantic similarity score into soft label, i.e., an initial rough estimation of the instance's usage (§ 4.2.2). Finally, we treat the soft labels as distant supervision for downstream probabilistic latent variable models, in which the usages are considered as the hidden variables and are represented over a set of features.

4.2.1 Literal Usage Representation

An idiom co-occurs with different sets of words depending on whether it is meant literally or figuratively. For example, when used literally, *get wind* is more likely to co-occur with words such as *rain*, *storm* or *weather*; in contrast, when used figuratively, it frequently co-occurs with *rumor* or *story*, etc. Comparing the two sets of words associated with the idiom, we see that the literal set of words also tend to co-occur with just *wind*, a constituent word within the idiom. Therefore, even without annotated data or dictionary, we may still approximate a representation for the literal meaning of an idiom by the idiom's constituent words and their semantic relationship to other words. To do so, we begin by initializing a *literal meaning set* to just the idiom's main constituent words[2]; we then grow the set by adding two types of semantically related words. First, we look for

[2] We observe that the nouns tend to be the most indicative of the idiom's literal meaning, but if the idiom does not contain any noun, we back off to any constituent word that is not a stop word.

co-occuring words in a large textual corpus (e.g., [David et al., 2005]): for each constituent word w, we randomly sample s sentences that contain w from the corpus; we extract the top n most frequent words (excluding stop words) and add them to the literal meaning set. Second, we look for words that are semantically close in a word embedding space: we train a continuous bag-of-words (CBOW) embedding model [Mikolov et al., 2013b] and add additional t words that are the most related to w using cosine similarity.

All together, the literal usage representation is a collection of vectors, i.e., the embeddings of the words in the final extended literal meaning set. The size of the set depends on parameters s, n, and t; if the chosen values are too small, we do not end up with a word collection that is representative enough; if the numbers are too large, we would only be wasting computing resources chasing Zipfian tails. Parameter setting choices are discussed further in the experiment section.

4.2.2 Literal Usage Metrics

Among all the instances to be classified, we expect the context words of the literal cases to be more semantically close to the literal usage representation we just formed. Let L denote the set of words in the literal usage representation for the target idiom. For each instance, let C be the set of non-stop context words in the instance. We calculate s, the semantic similarity score between the context of the instance and the literal usage representation as follows:

$$s = \frac{1}{|C|} \sum_{c \in C} \frac{1}{|L|} \sum_{l \in L} sim(c, l) \tag{4.1}$$

where c denotes a word in C, l denotes a word in L and $sim(c, l)$ refers to the cosine similarity between the word embeddings of c and l.

Let $S = \{s_1, s_2, ...s_n\}$ be the set of semantic similarity scores for all the instances we wish to classify. Instances with higher scores are more likely to use the idiom literally. A naive literal usage metrics is to choose a predefined threshold for all idioms and label all the instances with score above the threshold as literal usages. This approach is unlikely to work well in practice. As noted by previous work, idioms have different levels of **semantic analyzability** [Gibbs et al., 1989, Cacciari and Levorato, 1998]. When an idiom has a high degree of semantic analyzability, its

contextual words will be more semantically close to the literal usage representation, thus a higher threshold is needed.

In this work, we select a different decision threshold for each idiom adaptively based on the similarity scores distribution. And most importantly, rather than generate a hard label, we transform these scores into a probabilistic metric, where 0 means the usage in the instance is almost certainly figurative while 1.0 means it is literal.

We propose a metric based on the principle of **Minimum Variance (MinV)**. That is, we first sort the scores in S and choose the threshold (from these scores) that minimizes the sum of variances of the two resulting clusters. For each instance i, we then apply the following metric to estimate the probability that the idiom in instance i is meant **literally** based on its semantic similarity score s_i :

$$Pr_i = \frac{1}{1 + e^{-k*(s_i - t)}} \tag{4.2}$$

where k is a constant weighting factor and t indicates the learned threshold. The intuition is that the larger the difference between s_i and the threshold is, the more likely the instance i is literal; the probability of literal usage is not linearly correlated to the difference, we use the sigmoid function to account for this non-linearity. We incorporate *k* to scale the value of the difference since it is generally very small (close to 0). Without *k*, all the Pr values gravitate toward 0.5, rendering the soft label being equivalent to random guess. We set k to 5 for all the idioms based on a development set.

4.2.3 Heuristically Informed Usage Recognition

The soft label, generated by MinV (the literal usage metric), captures the distributional semantic information of the context. In practice, there are a variety of other linguistic features which are also informative of the intended usage of idiom. We explore probabilistic latent variable models over a collection of features that are linguistically relevant for idiom usage detection. The soft label is integrated into the unsupervised learning of hidden usages as a distant supervision. In this section, we will describe the proposed features in the latent variable models and how we integrate the soft label into the learning process.

4.2.3.1 Latent Variable Models To predict an idiom's usage in instances, we consider two representative probabilistic latent variable models: unsupervised Naive Bayes (NB) and Latent Dirichlet Allocation (LDA) [Blei et al., 2003][3] . For both models, the latent variable is the idiom usage (figurative vs. literal); the observables are linguistic features that can be extracted from the instances, described below:

Subordinate Clause We encode a binary feature indicating whether the target expression is followed by a subordinate clause (the Stanford Parser [Chen and Manning, 2014] is used). This feature is useful for some idioms such as *in the dark*. It usually suggests a figurative usage as in *You've kept us totally in the dark about what happened that night*.

Selectional Preference Violation of selectional preference is normally a signal of figurative usage (e.g., having an abstract entity as the subject of *play with fire*). We encode this feature if the head word of the idiom is a verb and focus on the subject of the verb. We apply Stanford Name Entity tagger [Finkel et al., 2005] with 3 classes ("Location", "Person", "Organization") on the sentence containing the idiom. If the subject is labeled as an Entity, its class will be encoded in the feature vector. Pronouns such as "I" and "he" also indicate the subject is a "Person". However, they are normally not tagged by Stanford Name Entity tagger. To overcome this issue, we add Part-of-Speech of the subject into the feature vector.

Abstractness Abstract words refer to things which are hard to perceive directly with our senses. Abstractness has been shown to be useful in the detection of metaphor, another type of figurative language [Turney et al., 2011]. A figurative usage of an idiomatic phrase may have relatively more abstract contextual words. For example, in the sentence *She has lived life in the fast lane*, the word *life* is considered as an abstract word. This is a useful indicator that *in the fast lane* is used figuratively. We use the MRC Psycholinguistic Database Machine Usable Dictionary [Coltheart, 1981] which contains a list of 4295 words with their abstractness measure between 100 and 700. We calculate the average abstractness score for all the contextual words (with stop words being removed) in the sentence containing the idiom. The score is then transformed into categorical feature to overcome sparsity problem based on the following criteria: concrete (450 - 700), medium (350 - 450), abstract (100 - 350).

[3] Although originally conceived for modeling document content, LDA can be applied to any kind of discrete input

Neighboring Words Words preceding and following the idiomatic expression can be very informative in terms of usage recognition. For example, words such as *relax* or *shower* before the idiom *in hot water* often signal a literal usage.

Part-of-Speech of the Neighboring Words Class of neighboring words might be useful as well. For example, a pronoun preceding *dog's age* generally indicates a literal usage, as in *I think my dog's age is starting to catch up. She sometimes needs help to jump on to my bed*, while a determiner usually marks a figurative usage, as in *It's been a dog's age since I've used Twitter.*

4.2.3.2 Incorporating Soft Label into Usage Recognition

Given a collection of instances and their features, either LDA or NB can separate the instances into two groups (hopefully, by usages), but it does not associate the right label (i.e., "figurative" or "literal") to the groups. We do not want to rely on any manual annotations for this step. Therefore, we integrate the automatically generated soft labels (based on MinV, our literal usage metric) into the unsupervised learning procedure as a weak form of supervision. Formally, we want to estimate each instance's posterior distribution over (literal/figurative) usages θ_{du} and usage-feature distribution ϕ_{uf}. For LDA, we derive a Gibbs sampling algorithm which incorporates the soft label into the learning procedure. We refer it as informed Gibbs sampling (infGibbs). For unsupervised naive Bayes model, we adapt the classical Expectation-Maximization algorithm to integrate the soft label. We refer it as informed Expectation-Maximization (infEM).

Informed Gibbs Sampling The Gibbs sampling algorithm [Griffiths and Steyvers, 2004] used in traditional LDA initializes each word token a random hidden topic. The system needs to interpret the learned topics post-hoc, e.g., by human annotation. In our case, for each feature f in each instance, an initial random usage **biased** by the instance's soft label is assigned to f (i.e., a Bernoulli trial). Since the soft label explicitly encodes an instance's literal and figurative usage distribution, we do not need to interpret the learned usages at the end of the algorithm. Based on these assignments, we build a feature-usage counting matrix C^{FU} and instance-usage counting matrix C^{DU} with dimensions $|F| \times 2$ and $|D| \times 2$ respectively ($|F|$ is the feature size and $|D|$ is the number of instances): $C^{FU}_{i,j}$ is the count of feature i assigned to usage j; $C^{DU}_{d,j}$ is the count of features assigned to usage j in instance d. Then for each feature f in each instance, we resample a new usage for f and matrices C^{FU} and C^{DU} will be updated accordingly. This step will be

repeated for T times. The resampling equation is:

$$p(u_i = j|u_{-i}, f) \propto p_j \cdot \frac{C^{f_i}_{-i,j}+\beta}{C^{(*)}_{-i,j}+|F|\beta} \cdot \frac{C^{d_i}_{-i,j}+\alpha}{C^{d_i}_{-i,*}+|U|\alpha} \tag{4.3}$$

where i indexes features in the instance d, j is an index into literal and figurative usages, $*$ indicates a summation over that dimension and $-$ means excluding the corresponding instance. The first factor p_j is the soft label encoding prior usage distribution. The second factor represents the probability of feature f under usage j ($C^{f_i}_{-i,j}$ is the count of the feature f assigned to usage j, excluding the current usage assignment u_i). The third factor represents the probability of usage j in the current instance ($C^{d_i}_{-i,j}$ is the count of linguistic features which are assigned to usage j in the current instance, excluding the current feature f). The value of $|U|$ is 2, representing the number of usages (i.e., figurative and literal). α and β are the hyper-parameters from the Dirichlet priors (we set both of them to 1). The core idea of Equation 4.3 is to integrate both distribution semantic information (soft label, the first factor) and linguistically motivated features (the second and third factors) into the inference procedure.

The matrices of C^{FU} and C^{DU} from the last $10\% * T$ iterations are averaged and then normalized to approximate the true usage-feature distribution ϕ_{uf} and instance-usage distribution θ_{du} respectively. The final result is determined by θ_{du}, i.e., assigning each instance with the usage of probability higher than 0.5. We do average to have a more stable result because an accidental bad sampling would affect our model negatively if we only use the C^{FU} and C^{DU} from the last iteration. This procedure is important for some idioms if their feature space is sparse. The iteration number T is set to 500 based on a development set.

Informed Expectation Maximization Combining a Naive Bayes classifier with the EM algorithm has been widely used in text classification and word sense disambiguation [Hristea, 2013, Nigam et al., 2000]. In our case, we want to construct a model to recover the missing literal and figurative labels of the instances of the target idiom. This section describes two extensions to the basic EM algorithm for idiom usage recognition. The extensions help improve parameter estimation by taking the automatically learned soft labels into consideration.

Our informed EM method extends a basic version for NB [Hristea, 2013], where the initial parameter values θ_{du} and ϕ_{uf} are chosen randomly. At each iteration, the E-step of the algorithm estimates the expectations of the missing values (i.e. the literal and figurative usage) given the latest

iteration of the model parameters; the M-step maximizes the likelihood of the model parameters using the previously-computed expectations of the missing values. As we've done with extending Gibbs sampling for LDA, we also perform two similar adaptations on conventional EM for NB to incorporate soft labels. First, we assign each instance an initial usage distribution θ_{du} directly using the soft label, and then initialize the usage-feature distribution ϕ_{uf} using these assignments. We refer it as informed initialization. Second, in the E-step, we multiply the expectation result of the basic EM with the soft label as the new expected usage for each instance (i.e., updating θ_{du}). The M-step is the same as basic EM to update the usage-feature distribution ϕ_{uf}.

4.3 Evaluation

To verify our hypothesis that using the semantic distance between context and idioms as distant supervision can help to reduce the need of human supervision, we conduct a comparative study to address three questions:

1. How effective is our overall approach? How does it compare against previous work?
2. How effective is our literal usage metric (i.e., MinV) compared to other heuristics?
3. How effective is our literal usage metric at informing downstream learning processes?

4.3.1 Experimental Setup

Models Our full unsupervised model first uses MinV to generate prior usage probability for each instance, which will then be integrated into the parameter estimation algorithms: informed Gibbs and informed EM, in the downstream hidden variable models. Therefore, we have two full models: MinV+infGibbs and MinV+infEM. We report the average performance of our models over 5 runs. Performing multiple runs is necessary because we have a sampling process. They are compared with three baseline unsupervised models: [Fazly et al., 2009], [Sporleder and Li, 2009] and [Li and Sporleder, 2009]; and two baseline supervised models: [Rajani et al., 2014] and the ensemble model we proposed in Chapter 3.

Parameter setting Recall that in order to build the literal usage representation of an idiom, we

need to sample s sentences that contain each constituent word w from an external corpus; extract from them the top n most frequently co-occurring words with w; then separately find t words that are semantically similar to w using word embeddings. To set parameters with values in reasonable ranges, we evaluated MinV on a small development set. We picked 10 idioms that are different from the evaluation set, scraped 50 instances from the web for each idiom, and labeled them ourselves. We find that s >= 100, n=10, and t=5 yield good results.

We use the gensim toolkit [Řehůřek and Sojka, 2010] and train our word embedding model using the continuous bag of word model on Text8 Corpus[4]. Negative sampling is applied as the training method; the min_count is set to 2. For the other parameters, we use the default settings in gensim.

Evaluative Data We compare all the methods using SemEval 2013 Task 5B corpus, which is used by prior supervised methods [Rajani et al., 2014], and verb–noun combination (VNC) dataset, which is used by a prior unsupervised method [Fazly et al., 2009]. However, there are some methods-datasets conflicts that have to be resolved. Because the idioms in the SemEval dataset are all in their canonical forms, and because the idioms are not restricted to the verb-noun combination, we cannot evaluate the method by [Fazly et al., 2009] on this dataset (as their method is tailored to verb-noun combination).

4.3.2 The Performance of Our Full Models

Table 12 shows the result of our models and the other comparative methods. Our proposed models show consistent performance across the two corpora, outperforming the unsupervised baselines from [Sporleder and Li, 2009], [Li and Sporleder, 2009] and the supervised model from [Rajani et al., 2014]. Moreover, there is no statistical significance in the F-score difference between our supervised ensemble model presented in the previous chapter and MinV +inf-Gibbs.

On the VNC corpus, our models have comparable average scores as that of [Fazly et al., 2009]; our scores are more stable across different idioms. While the method of Fazly et al. is nearly perfect for some idioms (0.98 on "take heart"), it performs poorly for others (e.g., 0.33 on "pull * leg"). Their algorithm has trouble with idioms whose canonical and non-canonical forms can

[4] From http://mattmahoney.net/dc/text8.zip

Table 8: The performances of different models. Avg. F_{fig} denotes average figurative F-score, Avg.Acc denotes average accuracy. We report the range in the parenthesis. * indicates the difference is significant with our MinV+ infGibbs model at the 95% confidence level. Since the method from [Fazly et al., 2009] restricted their experiment to VNC type, we only report their performance on the VNC corpus.

Type	Model	SemEval		VNC	
		Avg. F_{fig}	Avg.Acc	Avg. F_{fig}	Avg.Acc
Unsupervised	Sporleder & Li	0.58* (0.42 $\sim$ 0.72)	0.52*(0.32 $\sim$ 0.7)	0.61* (0.46 $\sim$ 0.73)	0.57*(0.41 $\sim$ 0.75)
	Li & Sporleder	0.64* (0.41 $\sim$ 0.76)	0.62*(0.43 $\sim$ 0.71)	0.67* (0.48 $\sim$ 0.77)	0.66*(0.52 $\sim$ 0.77)
	Fazly et al.	-	-	0.73 (0.33 $\sim$ 0.98)	0.74 (0.35 $\sim$ 0.98)
Supervised	Rajani et al.	0.71* (0.54 $\sim$ 0.83)	0.75(0.67 $\sim$ 0.81)	0.69* (0.49 $\sim$ 0.8)	0.7*(0.6 $\sim$ 0.79)
	Our Ensemble Model	0.77 (0.68 $\sim$ 0.85)	0.77(0.71 $\sim$ 0.85)	0.75 (0.65 $\sim$ 0.88)	0.75(0.67 $\sim$ 0.89)
Our Model	**MinV + infGibbs**	0.75 (0.64 $\sim$ 0.91)	0.74(0.63 $\sim$ 0.87)	0.73 (0.64 $\sim$ 0.86)	0.75(0.66 $\sim$ 0.83)
	MinV + infEM	0.73 (0.58 $\sim$ 0.88)	0.73(0.61 $\sim$ 0.85)	0.72 (0.62 $\sim$ 0.87)	0.72(0.6 $\sim$ 0.84)

appear frequently both in literal and figurative usages.

4.3.3 Effectiveness of MinV

The core of our approach is MinV, the literal usage metric we developed to generate soft labels to guide the unsupervised learning. This experiment examines its effectiveness by creating usage classifications directly from it (i.e., if MinV predicts a probability of $>$0.5, predict "literal"). We compare MinV against two alternative heuristics.

MinV is based on two core ideas. First, if an idiom is used figuratively, we expect to see a big difference (low similarity scores) between its context and the semantic representation of idiom's literal usage. The idea is similar to that of [Sporleder and Li, 2009], but they relied on lexical chain instead of distributional semantics. Second, instead of choosing a predefined threshold to separate the raw semantic similarity scores, we select a different decision threshold for each idiom adaptively based on the distribution of the scores. So as an alternative, we compare MinV against a Fixed-Threshold heuristic that labels an instance as "literal" if its raw score is higher than some

Table 9: A comparison of classifying by different heuristics. Results are averaged across all the idioms in the two corpora.

Model	Avg. F_{fig}	Avg.Acc
Fixed-Threshold	0.6 (0.23 $\sim$ 0.82)	0.62 (0.47 $\sim$ 0.83)
MinV	0.66 (0.43 $\sim$ 0.88)	0.65 (0.51 $\sim$ 0.89)
Sporleder & Li	0.59 (0.42 $\sim$ 0.73)	0.54(0.32 $\sim$ 0.75)

global threshold (set to 0.346 based on development data).

In Table 9, we observe that Minv outperforms both Sporleder and Li's model as well as Fixed-Threshold, but using MinV by itself is not sufficient. It has great fluctuations, e.g., the F-Score for individual idioms varies from 0.43 to 0.88. Recall that MinV +infGibbs has a smaller fluctuation across different idioms in Table 12. These results suggest that the subsequent learning process is effective.

Through error analysis, we find two major factors contributing to the performance fluctuation. First, the context itself could be misleading. An error case of **play ball** by MinV is:

*All 10-year-old Minnie Cruttwell wants to do is play with the boys , but the **Football Association** are not playing ball. She is a **member** of a mixed **team** called Balham Blazers , but the FA say she must **join** a girls' **team** when she is 12.*

The context words in bold (which are related to the word "ball") mislead MinV to predict a "literal" usage when it is actually a "figurative" usage (since an organization such as the Football Association cannot literally *play ball*). Second, the inclusion of all the content words in the context is not intelligent enough; there are words that do not provide useful semantic information in terms of distinguishing literal and figurative usages. Pruning the contextual words intelligently might result in more reliable models. We will leave this as the future work.

Table 10: The performance of MinV+NN and models without soft label on all the idioms in the two corpora.

Model	Avg. F_{fig}	Avg.Acc
Gibbs	0.58 (0.31 $\sim$ 0.78)	0.57 (0.4 $\sim$ 0.78)
EM	0.56 (0.31 $\sim$ 0.71)	0.6 (0.42 $\sim$ 0.77)
MinV+NN	0.68 (0.41 $\sim$ 0.83)	0.67 (0.55 $\sim$ 0.86)

4.3.4 Integration of MinV into Learning

We have argued that an advantage of using a metric with a probabilistic interpretation instead of a binary class heuristic is that its scores can be incorporated into subsequent learning models as soft labels. In this set of experiments, we evaluate the impact of the metric on the learning methods. First, we consider unsupervised learning without input from the literal usage metric. We cluster the instances with the original Gibbs sampling and EM algorithms and then label the two clusters with the majority usage within the clusters. Second, we explore using the information from the literal usage metric as "noisy gold standard" to perform supervised training on a nearest neighbors (NN) classifier. Specifically, the literal and figurative instances labeled by MinV with high confidence (top 30$\%$) are used as example set. Then for each test instance, we calculate its cosine similarity (in feature space) to the literal and figurative example sets and assign the label of the closest set. We refer this model as MinV +NN.

Table 10 shows the performances of the new models, which are all worse than our full models MinV +infGibbs and MinV +infEM. This highlights the advantage of integrating distributional semantic information and local features into one single learning procedure. Without the informed prior (encoded by the soft labels), the Gibbs sampling and EM algorithms only seek to maximize the probability of the observed data, and may fail to learn the underlying usage structure.

The model MinV +NN is not as competitive as our full models. It is too sensitive to the selected instances. Even though the training examples are instances that MinV is the most confident about, there are still mislabelled instances. These "noisy training examples" would lead the NN

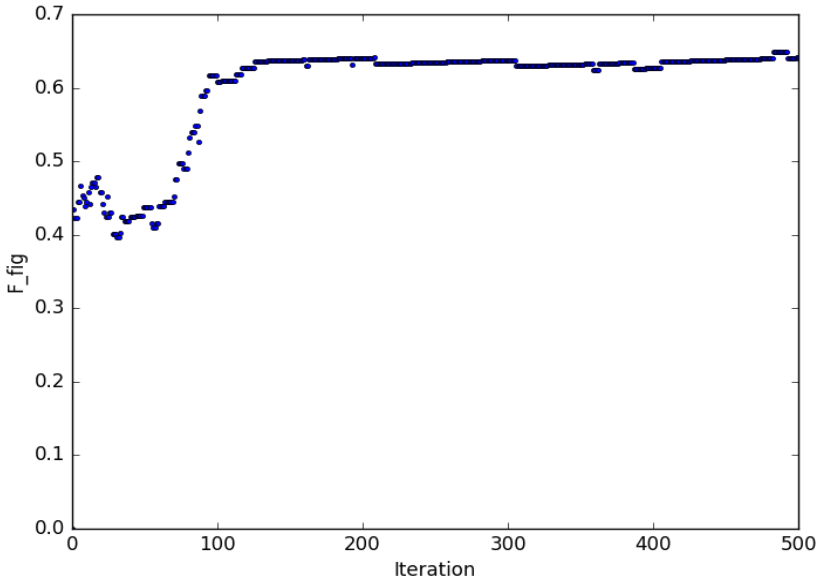


Figure 3: The performance of MinV+infGibbs on the idiom ”break a leg.”

classifier to make unreliable predictions. In contrast, our unsupervised learning is less sensitive to the performance of MinV; it can achieve a decent performance for an idiom even when the quality of the soft labels is poor. For example, when using MinV as a stand-alone model for *break a leg*, its figurative F-score is only 0.43, but through further training, the full model MinV+infGibbs achieves 0.64. Fig. 3 shows the training curve. A possible reason for this phenomenon is that the soft label is integrated into the learning process by biasing the sampling procedure (see Equation 3). We only encourage our model to follow the distributional semantic evidence captured by soft label and do not force it. So if there are strong evidences encoded by the linguistically motivated features in the instances to overcome the soft label it still has the freedom to do so. This is further supported by the fact that our full model MinV+infGibbs outperforms MinV on all the idioms.

4.3.5 Limitations

One limitation of this model concerns the literal usage representation. As we do not use any annotated labels, we approximate the literal usage representation by looking for words that are

associated with main constituent words of the idiom. The parameters in this process, such as the number of sentences sampled from Gigaword and the number of most frequently co-occurring words, are set based on a developing data and therefore subject to corpus bias. In addition, we mainly use nouns as the major constituent words. Although this strategy tends to work well on the two evaluative corpora, future studies should include more data to investigate whether it can be generalized to a broader range of idioms.

Another limitation of our model is that adapting it for resource-constrained languages would be challenging. The syntactic features (i.e., the subordinate clause feature) used in the latent variable models rely on robust syntactic parsers, which might be hard to get for low-resource languages. Moreover, we rely on the MRC Psycholinguistic Database Machine Usable Dictionary to measure the abstractness of the context. Due to the limited scope of application of abstractness measurement, other languages might do not have this type of resource.

4.4 Chapter Summary

We have presented an unsupervised method for idiom usage recognition. Our approach consists of two major parts. The first part defines a heuristic to predict the idiom's usage; the second part uses the heuristic to guide learning. Our heuristic is based on the idea that when an idiom is used literally, it should be semantically more similar to its context. Therefore, we come up with a representation for the idiom's literal semantics so that it can be compared with the context. In particular, we approximate the literal semantics with an aggregate group of words: words that are similar to the content constituents of the idiom, and words that are frequently associated with these constituents elsewhere. A second consideration in defining the heuristic is the score it outputs to predict the idiom's usage. Our solution is to consider all the raw similarity scores of instances for a target idiom as a population and scale them so that they take on a probabilistic interpretation. This allows the heuristic outputs to be used as soft labels that can then be integrated into a downstream probabilistic latent variable model, which can learn further without supervision to improve the final classification.

5.0 Generalized Idiom Usage Recognition via Semantic Compatibility

5.1 Introduction

To achieve our research goal, we have presented a supervised ensemble model to improve its robustness and an unsupervised model to reduce human effort. Although these two models have pushed the frontier of idiom usage recognition greatly, they are still computationally intensive, i.e., we need to collect a large number of instances for an individual idiom, either labeled or unlabeled, and train a specific model for that idiom. The abundance of idioms in text desperately calls for an efficient generalized model. However, the heterogeneity of idioms' behaviors makes a generalized model much more challenging than idiom specific models. Concretely, in the task of idiom usage recognition, different idioms could have varied context clues. For example, the idioms "play with fire" and "get wind" are used differently in the instances below. The proposition "of" following "get wind" often indicates the idiom is used figuratively (as in instance #4), while for idiom "play with fire", one might need more complicated linguistic clues to infer its usage, such as a violation of selectional preference (as in instance # 2).

(1) [*lit.*]Kids **playing with fire**: experts warn parents to look out for danger signs.

(2)[*fig.*]The UN is **playing with fire** over North Korea crisis.

(3)[*lit.*]Here in Portland we're just gonna get rain, the coast is gonna **get wind**. Stay safe!

(4)[*fig.*]FAA will **get wind** of that crooked airways' shady dealings.

Given these varied properties of idioms, it should be obvious that we cannot simply rely on superficial features (e.g., lexical feature) to build a generalized model. So alternatively, we resort to features that are invariant across idioms. The method we present in this section is based on the observation that when the literal interpretation of a potential idiomatic expression is not compatible with the context, it typically indicates that the idiom is used figuratively. For instance, in example #4 above, the word "wind" is generally far away from the surrounding words; the literal sense of "get wind" is not fit well with the context. In addition, early research work in psycholinguistics

also suggested that an idiom's figurative meaning will be retrieved from memory when the literal interpretation is rejected as it is not compatible with the context [A Bobrow and M Bell, 1973]. Generally, this *semantic incompatibility* is a strong indicator that the idiom has a non-literal interpretation in the context. It is at least possible, then, to build a generalized idiom usage recognizer by determining the semantic compatibility between the literal meanings of idioms and their contexts.

One way to measure semantic compatibility is with probabilistic language models (LMs), which assigns a probability to a sentence or a word sequence. In particular, n-gram with maximum likelihood estimation is often used. The most straightforward way to determine the intended usage of an idiom in an instance using n-gram is to calculate the probability of the instance. Intuitively, the figurative instances would have lower probability since the constituents of idiom are not compatible with the context; one can thus define a threshold for usage classification.However, it is hard to define a general threshold using n-gram because the probability of a target instance depends on the length of the sentence and the global frequencies of each word in it. Additionally, n-gram models are count based, which can not handle unseen combinations of tokens.

Alternatively, we find that the notion of semantic compatibility is reminiscent of the training objective of negative sampling in word2vec, which is originally used for learning low dimensional word embeddings [Mikolov et al., 2013b, Mikolov et al., 2013a]. Its Continuous Bag-of-Words (CBOW) variant internally tries to maximize the probability of positive(compatible) context-word pairs and minimize the probability of randomly sampled negative (incompatible) pairs. Thus if the CBOW can successfully capture the semantic compatibility feature in text, it is highly possible that we can apply it to determine the semantic compatibility between an idiom and its context.

However, the CBOW model mainly uses semantic compatibility as a roundabout way to learn useful vectors for words. The post-hoc evaluations of the model concentrate on the learned embeddings of words [Mikolov et al., 2013a, Levy et al., 2015], while whether the learned model can be directly applied to measure semantic compatibility is understudied. In this work, we analyze the potential limitations of the standard CBOW model in terms of semantic compatibility measurement (see Section 3.1). We further propose a novel semantic compatibility model by adapting the standard CBOW in two ways. First, we introduce several alternatives for context representation. We exploit bidirectional LSTM [Graves et al., 2013] to model the sequential information in con-

text and two self-attention mechanisms [Vaswani et al., 2017] to capture the critical context words. Second, we add a multilayer perceptron layer to relax CBOW's constraint on contextual similarity and tailor it for capturing semantic compatibility.

The overview of our method is shown in Figure 4. In our solution, the semantic compatibility model is used in a **transfer learning fashion**: (1) the model is first trained based on **large raw corpora** (such as Wikipedia) with the aim of predicting the semantic compatibility between context and **a single word**; (2) the learned model is then used to determine an idiom's intended usage by measuring the semantic compatibility between the idiom's literal sense and the context. Since idioms are multi-word expressions, we treat it as a single semantic unit and build a literal representation of idiom, which enables seamlessly reusing the semantic compatibility model for usage recognition. The advantages of our model are: (1) there is no need for annotated idiom usage examples since the core component of our usage recognition model (i.e., the semantic compatibility model) is trained on raw corpora; (2) the model is generalized, i.e., it can be applied to different idioms without further parameter tuning. We conduct experiments on two corpora; results suggest that the proposed generalized model achieves competitive results compared to state-of-the-art per-idiom models.

5.2 Background

Our model is built on the basis of CBOW and we explore attention mechanisms to capture the critical context words. Thus, we have a brief review of CBOW and attention mechanisms before proceeding to the model description.

5.2.1 Continuous Bag-of-Words

Neural language models which learn low dimensional vector word representation encoding both semantic and syntactic information, are currently one of the most influential methods in NLP [Bengio et al., 2003]. The probability of a sequence of words is calculated based on the learned vector representations, which can generalize well to unseen sequences of tokens. The word2vec

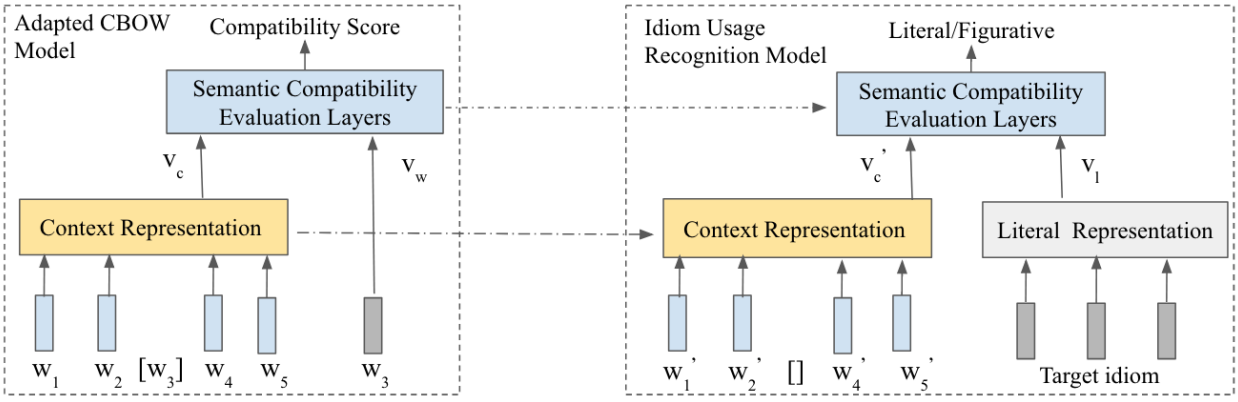


Figure 4: The overview of our idiom usage recognition model in a transfer learning fashion: the CBOW is adapted for semantic compatibility measurement which can be trained on raw large corpus; the learned representations and parameters are then used for idiom usage recognition. [] indicates target word or idiom.

[Mikolov et al., 2013a, Mikolov et al., 2013b] family of algorithms, developed from a shallow neural network, is an effective way to generate such embeddings.

The Continuous Bag-of-Words (CBOW) variant of word2vec internally tries to predict the target word based on the context words (as shown in Figure 5). Due to large vocabulary size, the training of CBOW is computational expensive. One widely applied strategy to speed up training is negative sampling. In particular, it defines two sets of embeddings: the "official" word embeddings and a second set of context embeddings, for each word in the vocabulary. The embeddings in the two sets are K-dimensional vectors which are tuned iteratively by scanning huge amounts of texts by a sliding window. For each observed pair of context and target word, the model samples several "negative" words which are not compatible with the context. The training objective is to maximize the probability of positive (compatible) context-word pairs and minimize the probability of negative (incompatible) pairs generated from a known noise distribution.

Specifically, the loss function used in CBOW is:

$$\log \sigma(v_c \cdot v_w) + \sum_{w_j \in W_{neg}} \log \sigma(-v_c \cdot v_{w_j}) \tag{5.1}$$

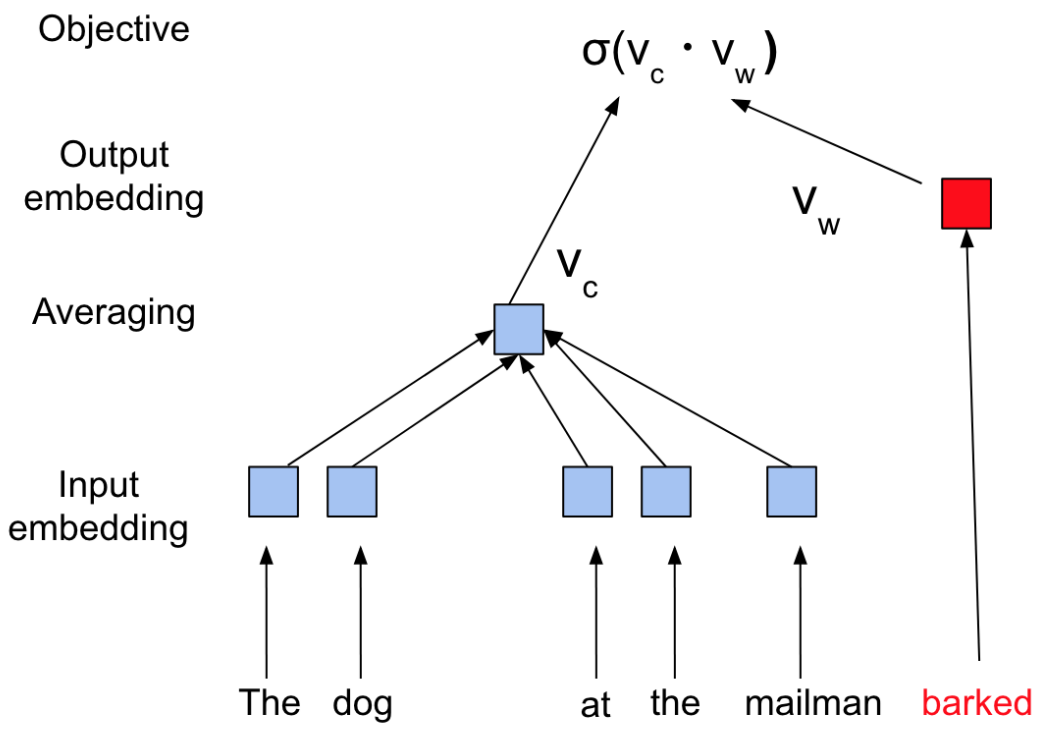


Figure 5: An working example of CBOW. Given the context "The dog () at the mailman", the model aims to assign those words which are fit to the context with high scores.

where v_c is the context embedding, v_w and v_{w_j} are the word embeddings of positive and negative target words, respectively. Since the sliding window usually contains more than one words, v_c is represented as the average of context embeddings of words within the window. The sigmoid function $\sigma(v_c \cdot v_w)$ can be considered as a semantic compatibility measurement; the model will update the context embeddings and word embeddings iteratively so as to assign high score to positive (compatible) pairs and lower score the negative (incompatible) pairs.

However, the goal of CBOW is limited to a smaller scope; its heavily trimmed network mainly aims to learn useful vectors for words to capture their semantic similarity. As semantic compatibility is essentially different with semantic similarity, we analyze the potential limitations of the standard CBOW model and adapt it to better model semantic compatibility in the following section.

5.2.2 Attention Mechanism

The fundamental task of neural networks is to allocate importance to input features through the weights of the neural network's model. In recent years, there has been a growing research interest in attention mechanism in deep learning community. Instead of using all available information, attention introduces a **memory-access** mechanism which can select the most important information in the learning process.

The application of attention in NLP is pioneered by [Bahdanau et al., 2014] in the task of machine translation. Before attention, state-of-the-art machine translation systems mainly used the Encoder-Decoder models. The Encoder first reads a complete sentence and compresses all information into one fixed-length vector. The Decoder then takes the vector as input and generates the translated sentence word by word. However, translation systems often deal with inputs and outputs of arbitrary length. A limitation of the Encoder-Decoder model is that it is has poor performance when translating long sentences since encoding long sentences using one single vector could lead to information loss. Attention mechanisms introduce a context vector in the Decoder. When generating a target word, the context vector is used to search for the relevant words in the source sentence. By utilizing this mechanism, it is possible for the Decoder to take into account the whole input to capture global information, rather than solely to infer based on one vector.

Apart from machine translation, attention has successfully been applied to tasks such sentence summarization [Rush et al., 2015] , question answering [Santos et al., 2016] and image captioning [Xu et al., 2015]. In these models, attentions have been typically used for alignment between two sources of information, e.g., the output and input sequences in machine translation [Bahdanau et al., 2014], or two input sequences such as question answering [Xiong et al., 2016, Lu et al., 2016].

For some tasks, however, there is no explicit alignment involved. For example, in sequence-to-one learning task, the input is just a single sequence of tokens; the model needs to relate different parts of the sequence in order to compute a representation of the same sequence for the purpose of classification. Researchers introduced self-attention (or intra-attention) to address this problem [Li et al., 2016, Lin et al., 2017, Cheng et al., 2016]. Notably, [Vaswani et al., 2017] showed that the self-attention could also be applied directly on raw word embeddings (without using sequence-

aligned recurrent architecture) for machine translation. The other applications of self-attention include question answering [Li et al., 2016] and sentiment analysis[Lin et al., 2017].

5.3 A Generalized Idiom Usage Recognition Model

We want to develop a generalized model for idiom usage recognition based on semantic compatibility. In this section, we first analyze the potential limitations of CBOW for semantic compatibility measurement. Then we present how we adapt the CBOW for semantic compatibility. Finally, we describe how we exploit the adapted model for idiom usage recognition.

5.3.1 Limitations of CBOW for Semantic Compatibility

As we have mentioned earlier, CBOW uses semantic compatibility as an auxiliary task to learn useful vectors for words to capture their similarity in hidden semantic space. An important question therefore is whether the learned context embeddings and the word embeddings, together with the sigmoid function, can be directly applied as a measurement of semantic compatibility. Although it seems plausible at first glance, we argue there are three potential limitations of CBOW that impedes it for semantic compatibility measurement.

1) A lack of sequential information To represent the context, CBOW simply uses the average of all the context embeddings, thus the order information is not preserved. Sequential models, such as the standard Recurrent Neural Network (RNN), can construct phrase and sentence representations in an order-sensitive way. They are becoming increasingly popular in NLP area because sequential information has been proved to be an important aspect for many applications such as text classification and sentiment analysis.

2) Not all words are equal In CBOW, all words contribute equally to the context representation. This limitation might not significantly impact the quality of the learned word embeddings, but could be problematic for semantic compatibility. In many cases, a few key context words are critical clues to determine the semantic compatibility between the context and a word.

3) A paradox of transitivity In CBOW, the **direct** dot product between context representation

and target word embedding is used to model their semantic compatibility. We find this dot product operation is not appropriate for encoding semantic compatibility relation; dot product aims to capture similarity relation ($\approx$) between two embeddings, which could lead to a paradox of transitivity in the case of semantic compatibility. In real world a word can appear in very different contexts. For example, in *John Lennon wrote a [song] called "Working Class Hero"* and *I like to listen to the same [song] on repeat*, the semantics of the two contexts of "song" are very different. Let C_1 and C_2 denote embeddings of two different contexts, i.e., $C_1 \not\approx C_2$. A target word B could be compatible with both $C1$ and $C2$ (as shown in the above example). If we use the direct dot product to model their compatibility, we can get $B \approx C_1$ and $B \approx C_2$ in the embedding space since B is compatible with both C_1 and C_2. Based on the transitive property of similarity relation, $C_1 \approx C_2$ can be inferred, which contradicts with the premise $C_1 \not\approx C_2$.

5.3.2 Adapting CBOW for Semantic Compatibility

We have discussed the potential limitations of CBOW for semantic compatibility. The first two limitations are related to context representations, while the third limitation is about the dot product operation. We propose to adapt the CBOW model to better capture semantic compatibility relation. In terms of context representation, we additionally use a special bidirectional Long Short-Term Memory network (LSTM) [Hochreiter and Schmidhuber, 1997] to encode sequence information. Meanwhile, we exploit self-attention mechanism [Lin et al., 2017, Vaswani et al., 2017, Li et al., 2016] to give more weight to important words when encoding context. Finally, instead of the simple dot product, a semantic evaluation layer is used to overcome the aforementioned paradox of transitivity.

5.3.2.1 Context Representation In standard CBOW, the context representation is the average of the embeddings of context words (denoted as ACE). Apart from ACE, we also exploit bidirectional LSTM for context representation, which has been shown effective for modelling sequential data [Graves et al., 2013, Melamud et al., 2016, Peters et al., 2018]. The overview of our architecture is illustrated in Fig. 6

Our architecture is not the same as standard Bidirectional LSTM [Graves et al., 2013]. In our

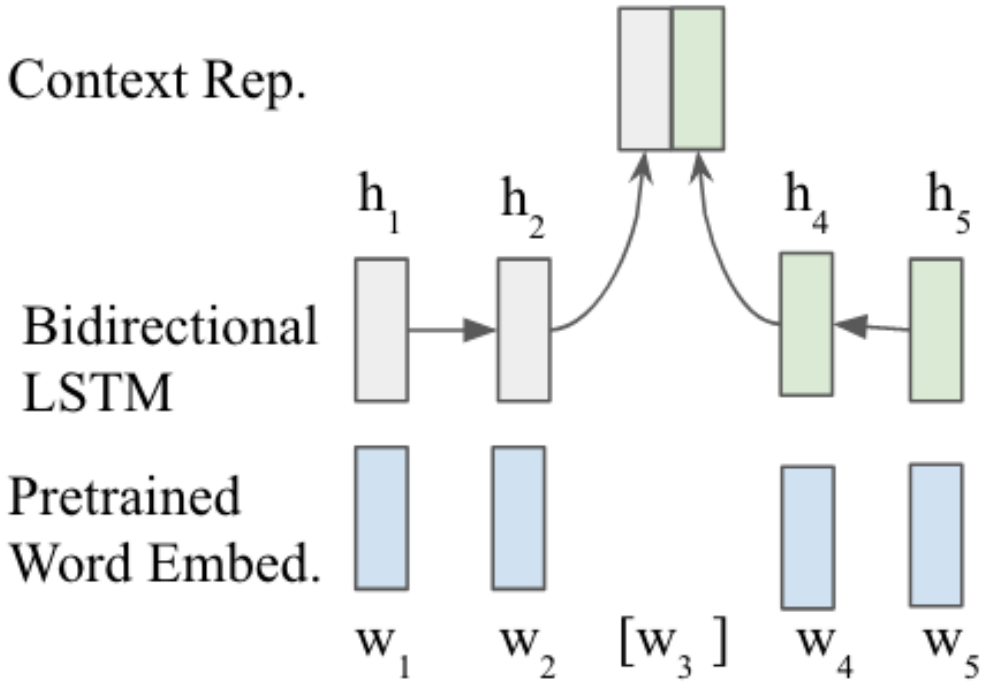


Figure 6: Bidirectional LSTM for context representation.

model, the two LSTMs gravitate toward the target words: a forward LSTM will generate a hidden representation for each word before the target word and a reversed LSTM will generate a hidden representation for each word following the target word; we do not feed the LSTMs with the target word itself. Let h be the hidden representation of word w (i.e., the output of the LSTMs), the context representation of the target word at position i is the concatenation of the hidden representations of the two neighboring words, i.e.,

$$c_i = [h_{i-1}; h_{i+1}] \tag{5.2}$$

Attention Layer In both ACE or the LSTM based context representation, we do not explicitly consider the importance of words. In this work, we exploit attention mechanism to enable our model to automatically identify those important words for semantic compatibility.

Attention mechanisms have generally been used to allow for an alignment of the input and output sequence, e.g. the source and target sentence in machine translation [Bahdanau et al., 2014], or for an alignment between two input sentences as in question answering [Santos et al., 2016,

Xiong et al., 2016]. In our work, we apply the idea of attention to a rather different kind of scenario, in which we only have the raw input sentence. We propose two self-attention (or intra-attention) models: global attention and local attention. The first one uses a vector to capture all the words that are important globally. As semantic compatibility usually involves the local interaction between words, our second attention model captures those words that have strong semantic relation with the other words in the context.

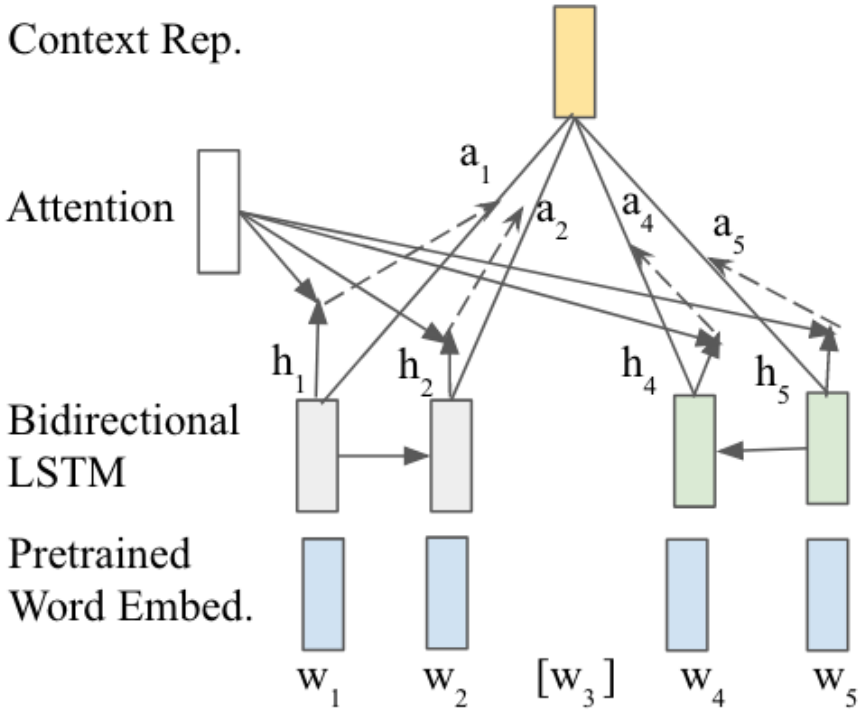


Figure 7: The global attention architecture when using bidirectional LSTM for sequential encoding.

Global Attention Figure 7 illustrates the global attention architecture when using bidirectional LSTM for context encoding. Assume v is the attention vector. The attention layer will generate an importance score g_i for each word w_i based on the dot product between v and its hidden representation h_i:

$$g_i = v \cdot h_i + b \tag{5.3}$$

Here the attention vector v is a parameter to be learned in the training process, which can be

considered as a global variable trying to "memorize" those critical words in a sentence based on the current context. The importance score is then normalized using softmax:

$$a_i = \frac{e^{g_i}}{\sum_{p=1}^{n} e^{g_p}}. \tag{5.4}$$

The attention-based context representation is a weighted sum of hidden states of LSTMs:

$$v_c = \sum_{i=1}^{n} h_i a_i. \tag{5.5}$$

Note that this global attention models can also be applied to the ACE for context representation. The only difference is the input to the attention layer: we only need to replace h_i in Equation 5.3 and 5.5 with the word embedding w_i.

Local Attention while global attention is useful, we argue it might not fully capture the semantic compatibility information in a sentence. A word that is important for semantic compatibility globally or in other sentences might not be important for the target sentence. Semantic compatibility usually involves the interactions among words within the sentence. We introduce a diagonal relevance matrix A with values $A_{i,j} = f(w_i, w_j)$ to characterize the strength of semantic interaction between words w_i and w_j. The scoring function f is computed as the inner product between the embeddings of w_i and w_j. If a word has strong semantic relation with another word, it is highly possible that this word is important. So we apply a max operation over the row of A (excluding the value in the diagonal because it is the relevance score between a word and itself) to select the largest value as the importance score for each word, i.e., ,

$$l_i = \max_j A_{i,j} \tag{5.6}$$

Following the global attention, a softmax layer is applied to normalize the raw score l_i; the final context representation is a weighted sum of hidden states of LSTMs. The overview of local attention is illustrated in Figure 8. Similarly, when applying local attention to ACE, the final context representation is a weighted sum of word embeddings.

5.3.2.2 Semantic Compatibility Evaluation Layer

To quantify the semantic compatibility between a context and a target word, standard CBOW uses the direct dot product between context

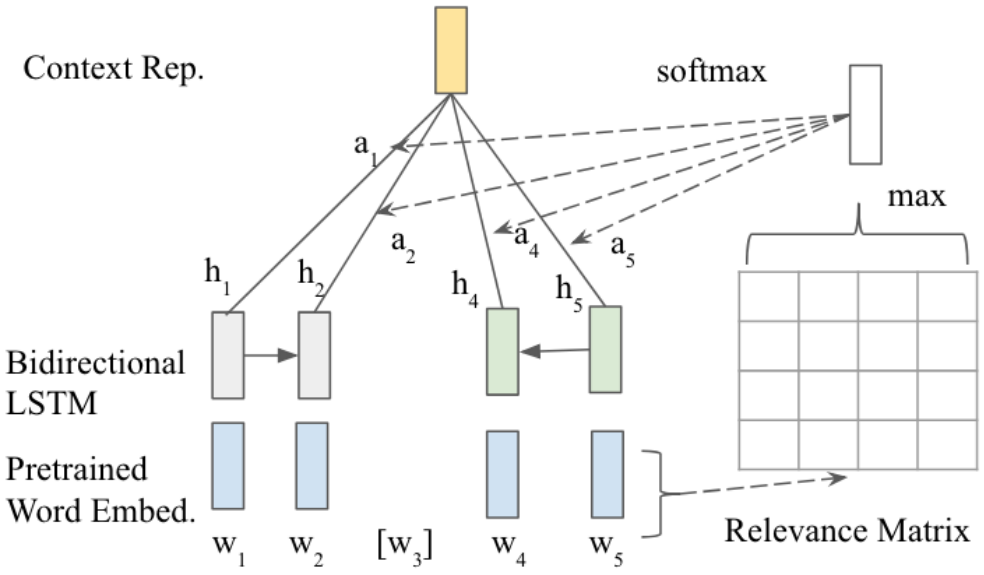


Figure 8: The local attention architecture when using bidirectional LSTM for sequential encoding.

embedding and target word embedding as the metric. We argue the direct dot product operation could lead to paradox of transitivity. To address this limitation, we apply a multilayer perceptron (MLP) with a ReLU nonlinearity over the context representation. The MLP is shown below in which the f_1 and f_2 denote fully connected layer.

$$\mathrm{L}(v_c) = f_2(relu(f_1(v_c))) \tag{5.7}$$

We use the following formula to measure the semantic compatibility between a context and a word:

$$\sigma(L(v_c) \cdot v_l) \tag{5.8}$$

Recall the main reason of paradox of transitivity is that a word can appear in very different contexts; the direct dot product between word embedding and context representation would, however, force these different contexts being similar to each other. This paradox is avoid by the multilayer perceptron L since it relaxes the contextual similarity constraints, i.e., it can map the context representations which are different originally to similar embeddings which are close to the target word. We refer the whole mapping and measuring schema as the semantic compatibility evaluation layer.

5.3.2.3 Training We train our adapted CBOW on wikipedia corpus using negative sampling. The loss function is:

$$\log \sigma(L(v_c) \cdot v_w) + \sum_{w_j \in W_{neg}} \log \sigma(-L(v_c) \cdot v_{w_j}) \tag{5.9}$$

The model is trained end-to-end using the Adam optimizer [Kingma and Ba, 2015]. Standard CBOW scans the whole corpus using a sliding window of a fixed size. Alternatively, we train the model sentence by sentence because using all the context words in a sentence can yield more precise context representation, which is essential for semantic compatibility.

5.3.3 Idiom Usage Recognition based on Semantic Compatibility

We have introduced how we adapt the standard CBOW for semantic compatibility measurement and train it on large corpus. Given a context representation and a word embedding, the learned model is expected to tell us whether they are compatible. However, we want to measure the semantic compatibility between a context and an idiom, which is usually a multi-word expression. To reuse the learned model, we first build a representation of the literal sense of the idiom. Then we use the semantic compatibility layer to evaluate whether the literal representation is compatible with the context.

5.3.3.1 Literal Representation of Idiom We experiment with the following two representations of the literal sense of idiom:

AWE, the average of the embeddings of words forming the idiom. The intuition is that the literal sense of idiom is compositional.

AKWE, the average of the embeddings of keywords in the idiom. Recall that when we built the literal usage representation in our unsupervised model (§ 4.2.1), we did not use all the constituents of the idiom. The intuition is that one or two words in idiom will be the crucial clue that indicates whether a figurative or literal sense was intended. Consider the figurative example of "get wind" at the beginning of this chapter, the word "wind" does not fit well to the context and this incompatibility servers as a strong signal of the intended usage, while the word "get" provides less information. In this work, for verb-noun combination, we only choose the noun as the keyword;

for noun-noun combination, we choose both nouns as the keywords; for the other types of idiom, the non-stop words are selected as the keywords. Although this representation might lose partial information of the literal interpretation of idiom, we hypothesize it could benefit our task.

5.3.3.2 Usage Classification Given a context representation v_c and the literal representation of idiom v_l, we calculate their compatibility score using the following formula:

$$\sigma(L(v_c) \cdot v_l + b_u) \tag{5.10}$$

where b_u is a bias term, which is tuned based on a development dataset. If the score is larger than 0.5, the instance will be classified as literal usage. Otherwise, it will be labeled as figurative usage.

5.4 Evaluation

To verify our hypothesis that using the semantic compatibility between contexts and idioms can help to train a generalized model, we conduct experiments to address the following questions:

1. How effective is our overall approach? How does it compare against previous work, especially the per-idiom models?

2. How effective is the standard CBOW for idiom usage recognition?

3. Does our model effectively address the limitations of CBOW?

5.4.1 Experimental Setup

Baselines We compare our models with four unsupervised models: [Sporleder and Li, 2009], [Li and Sporleder, 2009], [Fazly et al., 2009] and the model we presented in Chapter 4. For supervised model, we compare our models with [Rajani et al., 2014] and the ensemble model we presented in Chapter 3. All these models are per-idiom models except the one presented in [Sporleder and Li, 2009].

Our models We experiment with two base context representations: ACE and bidirectional LSTM, over which we additionally propose two attention models: local and global attention. Therefore

we have four variants for context representations. In terms of the representation of literal sense of idiom, we experiment with AWE and AKWE. So our full models have eight variants.

Parameter setting To train the adapted CBOW, we follow the standard training procedure in word2vec using negative sampling. To increase the training speed, we uniformly sampled a set of sentences from the Wikipedia [1] to build a corpus of 100M tokens. We find using corpus of this size is sufficient to train a reliable model so we do not use the full corpus. All those tokens with frequency less than 50 are trimmed. The hyperparameters are summarized in Table 11.

When applying the adapted CBOW model to idiom usage recognition, we need to set the bias term b_u in Equation 5.10 with value in a reasonable range. We picked 10 idioms that are different from the evaluation set, collected 50 instances from the web for each idiom, and labeled them ourselves. We find that b_u in the range of [0.06, 0.15] yield good results.

Table 11: Hyperparameters of our network.

Parameter	Value
word embedding size	200
context embedding size	200
LSTM hidden size	200
f_1 input/output size	200/400
f_2 input/output size	400/200
negative samples	15
epoch	10
batch size	500
learning rate	0.001

Following the experiment presented in Chapter 4, we compare all the methods using SemEval 2013 Task 5B corpus and Verb-Noun Combination (VNC) dataset.

[1] https://dumps.wikimedia.org/

5.4.2 Experimental Result

The result is shown in Table 12. We can observe that ACE+LocalAtt+AKWE gets an F-score of 0.76 (accuracy of 0.75) on SemEval corpus and 0.75 (accuracy of 0.73) on VNC corpus, which outperforms the per-idiom models from [Rajani et al., 2014], [Li and Sporleder, 2009] and the generalized model from [Sporleder and Li, 2009]. Moreover, the model is competitive to our ensemble model presented in Chapter 3.

Table 12: The performances of different models. Avg. F_{fig} denotes average figurative F-score, Avg.Acc denotes average accuracy. * indicates the difference is significant with our model ACE+LocalAtt+AKWE at the 95% confidence level. Since the method from [Fazly et al., 2009] restricted their experiment to VNC type, we only report their performance on the VNC corpus.

Type	Model	SemEval		VNC	
		Avg. F_{fig}	Avg.Acc	Avg. F_{fig}	Avg.Acc
Per-Idiom	Rajani et al., 2014	0.71*	0.75	0.69*	0.7
	Li and Sporleder, 2009	0.64*	0.62*	0.67*	0.66*
	Fazly et al., 2009	-	-	0.73	0.74
	Our Ensemble Model	0.77	0.77	0.75	0.75
	Our Unsupervised Model	0.75	0.74	0.73	0.75
Generalized	Sporleder & Li	0.58*	0.52*	0.61*	0.57*
Our Model	ACE + GlobalAtt + AWE	0.72	0.69	0.71	0.7
	ACE + GlobalAtt + AKWE	0.74	0.7	0.73	0.7
	ACE + LocalAtt + AWE	0.74	0.73	**0.76**	**0.73**
	ACE + LocalAtt + AKWE	**0.76**	**0.75**	0.75	**0.73**
	Bidirectional LSTM + GlobalAtt + AWE	0.68	0.68	0.67	0.67
	Bidirectional LSTM + GlobalAtt + AKWE	0.72	0.72	0.69	0.7
	Bidirectional LSTM + LocalAtt + AWE	0.69	0.68	0.7	0.69
	Bidirectional LSTM + LocalAtt + AKWE	0.73	0.72	0.72	0.71

5.4.3 Detailed Analysis

5.4.3.1 Using Standard CBOW for Idiom Usage Recognition In this study, we experiment using standard CBOW for idiom usage recognition, in which ACE is used as the context representation and the direct dot product between context representation and target word representation is used as a measurement of semantic compatibility. The training and evaluation procedures are the same as our full models.

Table 13: The results of CBOW for idiom usage recognition. Results are averaged across all the idioms in the two corpora.

Model	Avg. F_{fig}	Avg.Acc
CBOW+AWE	0.63	0.62
CBOW+AKWE	0.65	0.63

Table 13 shows the performance of CBOW for idiom usage recognition, which is significantly worse than our adapted models. Arguably, CBOW is insufficient to capture the semantic compatibility information in text. To illustrate this point, we compare the CBOW and our adapted model (we use the bidirectional LSTM + Local Attention for context representation) to select the most compatible words based on a given context. We find the results of CBOW remains of wildly-vary quality. Considering the example "can you see the [] i try to make?", the top 10 most compatible words to fill in the bracket predicted by the two models are shown in Table 14.

Table 14: Top 10 most compatible words in "can you see the [] i try to make?"

CBOW	Adapted CBOW
please	stuff
want	positives
you	ripples
hear	ones
how	things
try	changes
sure	figures
wish	pictures
know	dilema
do	negatives

As we can see, CBOW has a fairly poor semantic compatibility measurement; all the words

tend to make little sense in the context. In contrast, the adapted model has much better results. Since our idiom usage recognition heavily relies on the underlying model's ability of measuring semantic compatibility, this could potentially explain why the CBOW has a worse performance in the downstream task.

To better understand the effectiveness of sequential information, attention mechanism and semantic compatibility layer, we did an ablation study and the results are shown in Table 15. Since AKEW tend to outperform AWE (as shown in Table 12) , we only experiment with AKEW as the literal representation of idiom.

Table 15: The results of ablation study. Results are averaged across all the idioms in the two corpora.

Model	Avg. F_{fig}	Avg.Acc
ACE+GlobalAtt+AKEW	0.74	0.7
- w/o Semantic Layer	0.66	0.64
ACE+LocalAtt+AKEW	0.76	0.74
- w/o Semantic Layer	0.67	0.66
- w/o attention	0.66	0.67
Bidirectional LSTM+GlobalAtt+AKEW	0.71	0.71
- w/o Semantic Layer	0.65	0.64
Bidirectional LSTM+LocalAtt+AKEW	0.73	0.72
- w/o Semantic Layer	0.66	0.66
- w/o attention	0.7	0.69

5.4.3.2 Sequential Information The importance of sequential information is closely related to attention model. In Table 12, we can observe that our full non-sequential models (ACE variants) generally outperform the sequential models (Bidirectional LSTM variants). Without attention, however, we find sequential information can significantly boost the performance of our model; the bidirectional LSTM + AKEW achieves F-score of 0.7 while the ACE + AKEW only gets 0.66

as shown in Table 15. Intuitively, with the aid of attention, our model can identify those critical words, which enhances the expressiveness of context representation by simple weighted averaging.

5.4.3.3 Attention In Table 15, we can observe removing attention layer can result in performance drop for both ACE and bidirectional LSTM variants. This shows the effectiveness of our attention model in terms of context representation. Moreover, the global attention is not as competitive as the local attention. For example, the Bidirectional LSTM+LocalAtt+AKEW model achieves an averaged F-score of 0.73 on the two corpora while the Bidirectional LSTM+GlobalAtt+AKEW model gets 0.71. This observation aligns with our intuition that semantic compatibility usually involves the local interactions among words within the sentence. In Figure 9 we visualize the attention layer using the first example in the Introduction section. The global attention tends to assign higher weights to non-stop words such as "kids", "experts" and "sign", while the local attention tends to assign higher weights to words with strong semantic relation, such as "warn" and "danger".

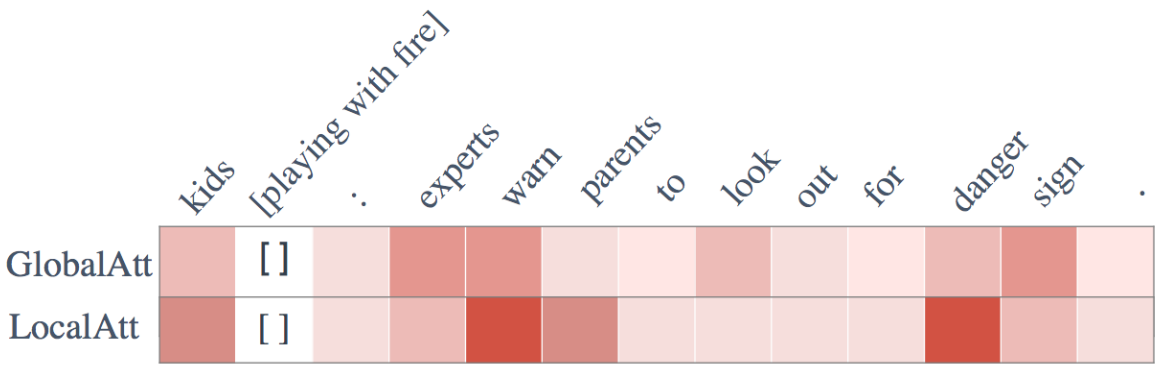


Figure 9: Visualization of attention layer.

5.4.3.4 The Semantic Compatibility Layer We have argued that the direct dot product between context representation and target word embedding could lead to the paradox of transitivity. To address this problem, we add a multilayer perceptron over the context representation so as to map different contexts to embeddings that are close to the target word.

In Table 15, we observe that the performances of our modes decrease significantly without the semantic compatibility layer. Among all the full models, the ACE+LocalAtt+AKEW has the most severe performance drop, i.e., from 0.76 to 0.67 in terms of F-score and 0.74 to 0.66 in terms of accuracy. This suggests the semantic compatibility layer is essential to our model.

5.4.4 Limitations

We have used ablation studies and visualization to demonstrate the effectiveness of the proposed attention models. But does our model always successfully capture the important words in context for semantic compatibility? To make the study more reliable, we can manually annotate those critical words and check whether they align with the weights learned by the attention models. However, human annotations require a great amount of time and effort outside the scope of this thesis. Alternatively, we can use statistical weighting methods such as tf-idf or syntactic parser to automatically label some candidate words and compare them with the attention models. A potential drawback of this method is that the candidates are produced using general-purpose weighting methods, which might deviate from actual critical words for semantic compatibility.

Another limitation concerns the evaluation of semantic compatibility. Although the ablation study suggests that the semantic compatibility layer is essential to our full models, a systematic quantitative evaluation is needed to justify that our models have learned the knowledge of semantic compatibility. As we can see from Table 14, we still have some predicted words that are not compatible with the context. In the future, we plan to use sentence completion task to evaluate our semantic compatibility model. In corpora such as Microsoft Sentence Completion Challenge (MSCC) [Zweig and Burges, 2011], each entry is a sentence with one word replaced by a gap. The task is to choose a word, out of five choices, that is most coherent to fill the gap. However, it is worth noting that general sentence completion tasks might need domain knowledge, reasoning and grammar analysis; sentence compatibility might get involved in only a small portion of the relevant datasets (e.g., MSCC).

5.5 Chapter Summary

To reduce the computational cost, we have built a generalized idiom usage recognition model such that it no longer needs to be trained separately for each individual idiom. Our idea is to quantitatively measure the semantic compatibility between the literal meanings of idioms and their contexts and use the result to determine the usages of idioms. Although the concept of semantic compatibility is reminiscent of the training objective of CBOW, we find that the standard CBOW can not fully capture the semantic compatibility in text due to its shallow architecture. We have developed a novel semantic compatibility model by addressing the limitations of the standard CBOW for the purpose of idiom usage recognition. Experiments have shown that the proposed generalized model achieves competitive results compared to the per-idiom models.

6.0 Applications of Idiom Usage Recognition Models

The ubiquity of idiomatic expressions in different genres of text has negative impacts on many NLP applications due to their idiosyncratic behavior. Recently, automated processing of idioms has been actively investigated to mitigate such impact [Liu et al., 2017, Cap et al., 2015, Fadaee et al., 2018, Spasic et al., 2017, Williams et al., 2015]. The models we have proposed in this thesis can be extended into semantic-related NLP tasks to address the ambiguity problem introduced by idioms. In this chapter, we first briefly discuss some potential applications of our models. Then we present a case study in which we integrate our models into modern machine translation system to improve its performance on sentences containing idioms.

6.1 Potential Applications

Information Retrieval Lexical ambiguity is a long-lasting problem for advanced information retrieval systems. For example, when one aims to search for information of the ”Apple” company, the results which are related to the fruit ”apple” should be excluded. This problem also occurs to idioms. When an idiom is used figuratively, its constituents would not have their literal interpretations. Therefore, when we search for information which is related to the literal senses of these constituents, information retrieval systems should be able to exclude sentences in which the idiom is used figuratively. Our model can help information retrieval system to achieve such a goal.

Automated Essay Scoring Previous studies had shown that appropriate use of idioms is a strong indicator of the native-like proficiency of the language and might be a reliable measure of writing skills [Cowie et al., 1984]. Therefore, our model can potentially benefit automated essay scoring [Ong et al., 2014, Persing and Ng, 2015]. For example, our model can help locate the figurative usages of idioms in essays and this information can serve as features for downstream automated essay scoring models.

Sentiment Analysis Idioms are commonly used in reviews and comments because they typically imply an affective stance toward something (rather than a neutral one) [Williams et al., 2015,

Nunberg et al., 1994]. Since words are the basic sentiment units in modern sentiment analysis models, studies reveal that a large number of errors of sentiment classification are caused by idioms due to their non-compositional property [Balahur et al., 2013, Williams et al., 2015]. [Williams et al., 2015] has shown that the inclusion of idioms as features can improve the performance of traditional sentiment analysis. Since the sentiments of literal and figurative usages of idioms might be different, it is promising that the usage information can potentially further boost the sentiment analysis models.

Machine Translation As we have mentioned at the beginning of this thesis, machine translation has a poor performance on sentences with idioms due to the usage ambiguity; state-of-the-art machine translation models generally treat idioms as normal expressions and are not sophisticated enough to translate them properly in different context. How to integrate the information learned by the idiom usage recognizers into advanced machine translation models is an interesting question to answer.

6.2 Case Study: Improving Machine Translation of Idioms

The majority of previous work on idiom translation mainly augments machine translation models with features indicating whether there is an idiom in the source sentence [Fadaee et al., 2018, Salton et al., 2014]. In this case study, we investigate whether the usage information of idiom (extracted by our usage recognition model) can benefit machine translation on idiom translation.

6.2.1 Integrating Usage Information into Machine Translation Model

To conduct the study, an important challenge is to build a dedicated parallel corpus of reasonable size for learning and evaluating idiom translation. We find the English-German idiom corpus from [Fadaee et al., 2018] to satisfy our need. This corpus is built from the data used in the WMT German-English Shared Task from 2008 to 2016 [Bojar et al., 2017]. Specifically, we perform the English-to-German translation task and each English sentence in the testing data contains at least one idiom in the dict.cc online dictionary. The statistics of the dataset are listed in Table 16 .

Table 16: Statistics of English-to-German translation dataset.

Number of unique idioms	132
Training size	4.5M
Idiomatic sentences in training data	1998
Test size	1500

Another challenge of this study is to integrate our usage recognition model into modern machine translation models. The full pipeline has to address many problems. First, it needs to locate the potential idioms in the sentence. Second, it has to recognize the usages of the potential idioms. Finally, we need to find a way to encode the usage information into machine translation models. As we have addressed the second problem in the previous chapters (we use the generalized model in this study), we need to address the first and the third problem in this study.

For each sentence in the English-to-German translation dataset, the idiom information (e.g., whether there is an idiom and the standard form of the idiom) is provided; we only need to find the position of the given idiom. We employ lexico-syntactic patterns to recognize their occurrences. Specifically, we first use exact string matching to locate them in text. It cannot find all the idioms since many idioms can also undergo certain syntactic changes such as inflection. To resolve this problem, we further use regular expressions to recognize their occurrence. To encode the usage information into machine translation models, a straightforward method is to append a special extra token $< fig >$ to each source sentence containing a figurative usage of idiom. This simple approach tends to be effective in machine translation systems which employ sequence-to-sequence architectures [Fadaee et al., 2018]. As this method ignores the position of the idiom, we also experiment with another method in which we insert a token $< start\_fig >$ before the idiom and a token $< end\_fig >$ after the idiom. We compare these two methods with the conventional setting in which no extra information regarding the usage of idiom is provided.

We use OpenNMT [Klein et al.,] to implement the machine translation model. The NMT

vocabulary is limited to the top 20K most frequent words in both languages. The hyperparameters are summarized in the following tables:

Table 17: Hyperparameters of our machine translation model.

Parameter	Value
Encoder layer	4
Encoder LSTM hidden state size	1000
Dropout	0.1
Epoch	20
Batch size	100

We use BLEU to measure the quality of translations. From the result presented in Table 18, we can see that the baseline achieves a BLEU score of 17.2, which is lower than the performance of previously reported models on the standard test set (WMT 2008-2016) [Sennrich et al., 2016]. This suggests that it is much harder to translate sentences containing idioms. Further, simply appending the $< fig >$ token to indicate the usage of idiom gets a BLEU score of 16.6, which is slightly lower than the baseline model; using the $< start\_fig >$ and $< end\_fig >$ tokens outperforms the baseline by 2.3 BLEU. This suggests that the usage information and the position information of the idiom can help boost the performance of neural machine translation models on idioms.

Table 18: The performance on English-to-German idiom translation test set.

Model	BLEU
NMT Baseline	17.2
with $< fig >$ token	16.6
with $< start_fig > < end_fig >$token	19.5

6.2.2 Limitations

As we have mentioned above, the idiom information is provided for each sentence in our study. In real application, however, we need to know whether there is an idiom in a sentence in the first place. One straightforward way is to rely on external idiom resources. For example, we can first build an up-to-date idiom dictionary of broad coverage and high quality (online dictionaries such as thefreedictionary.com and dict.cc are reasonable choices) and then use lexico-syntactic patterns to recognize whether an idiom in the dictionary occurs in the sentence. When the external idiom resources are not available, we can alternatively resort to idiom type classification methods to find potential idioms in a sentence [Fazly and Stevenson, 2006, Venkatapathy and Joshi, 2005, Katz and Giesbrecht, 2006].

Another concern is related to the figurative meanings of idioms. We only integrate the usage and position information of an idiom into machine translation models. Thus, we expect the models can learn the figurative interpretation of idioms from the training data. This is problematic for idioms with low semantic analyzability, especially when they do not have enough figurative instances for training. One solution to address this problem is to replace idioms with their figurative meanings in literal English. We have discussed this solution in [Liu and Hwa, 2016] and we will leave this as future work.

7.0 Conclusion

7.1 Summary

In this thesis, we have investigated how to build robust and efficient idiom usage recognizers so that the models can be applied to a broader range of idioms. We have hypothesized that our goals can be achieved through better modeling the interaction between idiom and context (§ 1.2). In Chapter 3, we have proposed an ensemble model which can draw knowledge from different representations. Experiment result (§ 3.4) supports the first hypothesis of this thesis that a robust idiom usage recognizer can be trained by addressing the interaction between context representations and linguistic properties of idioms (**H1** in § 1.2). In Chapter 4, we have presented an unsupervised idiom usage recognizer to reduce human effort. The competing performance (§ 4.3) of this unsupervised model supports the second hypothesis of this thesis that the semantic similarity between context and idiom can be used as distant supervision (**H2** in § 1.2). In Chapter 5, we have presented a generalized idiom usage recognition model by evaluating the semantic compatibility between context and the literal sense of the idiom. The generalized model can reduce the computational cost because there is no need to train the model for each individual idiom. This supports the third hypothesis of this thesis (**H3** in § 1.2). To demonstrate the application of our model, we have conducted a study in which we integrate the usage information of idioms into machine translation systems (§ 6.2). The following is a summary of our contribution.

- We have conducted the first study that analyzes the impact of linguistic properties of idioms on the effectiveness of context representations. Concretely, we focused on the semantic analyzability and context diversity of idioms. We have defined two metrics to quantitatively analyze their interactions with different representations of context.
- We have presented a supervised ensemble approach to adaptively combine multiple contextual semantic representations for different idioms. Our model can achieve better stability without loss of accuracy.
- We have proposed a novel *literal usage metric* based on the semantic similarity between the context and the idiom to estimate the likelihood that the idiom is used literally.

- We have shown how to use two representative probabilistic latent variable models (i.e., Latent Dirichlet Allocation and Naive Bayes) for unsupervised idiom usage recognition. The usage of an idiom is considered as the hidden variables and represented as a mixture of linguistically motivated features.
- We transformed the proposed *literal usage metric* into soft labels; we have further presented learning algorithms in which the soft label was served as distant supervision to guild the down-stream probabilistic latent variable models to better infer the usages of idioms. Our full model is competitive against supervised methods.
- We have presented a transferred learning approach for developing a generalized idiom usage recognizer. The model was trained on a large raw corpus and there is no need to annotate idiom usage examples for training.
- We have introduced the concept of semantic compatibility and proposed a novel semantic compatibility model by adapting the training of the Continuous Bag-of-Words (CBOW) model.
- We have successfully applied the semantic compatibility model on idiom usage recognition by measuring whether the literal senses of idioms are compatible with the contexts. Results have shown that our method achieves competitive results compared to state-of-the-art per-idiom models.
- We have presented a simple approach to extend our models into modern machine translation model. Results have shown that our models can improve the translation quality of idioms in text.

7.2 Future Work

The findings reported in this thesis open the door for a variety of future work. We discuss below some short term future work (§ 7.2.1) and open research questions (§ 7.2.2).

7.2.1 Short Term Future Work

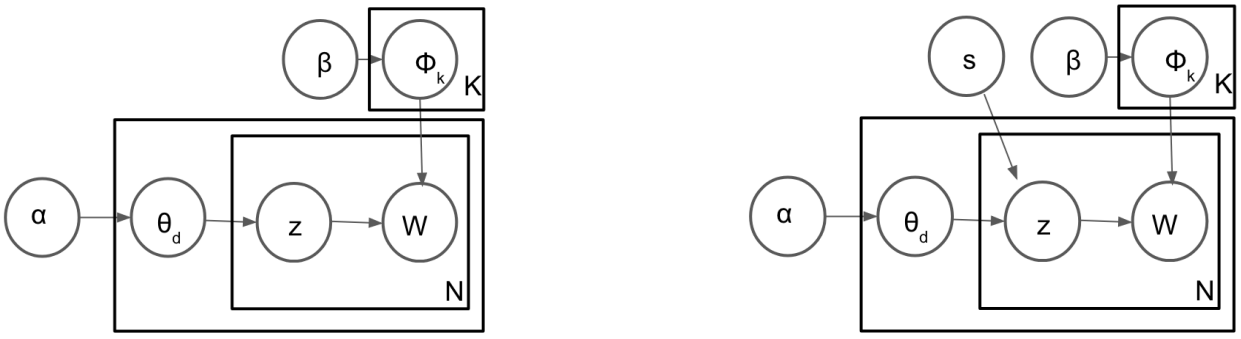
Weakly Informed Unsupervised Learning Unsupervised learning is one of the most active and productive areas in recent years in NLP. In Chapter 4, we generate soft labels encoding the likelihood of usages of idioms and use it as a form of distant supervision for downstream unsupervised models. We have empirically shown that the soft labels not only provide good initialization for the subsequent unsupervised methods, but also effectively guide the models toward grouping by usages. This unsupervised learning framework extends far beyond idiom usage recognition.

We propose to investigate the weakly informed topic modeling by adapting the standard LDA using the framework established in § 4.2.3 (as shown in Fig.10). Although the LDA model can infer topics based on given training documents, it does not associate the right labels (i.e., "politics" or "sport") to the topics; we need to manually interpret the learned topics after training. Our unsupervised learning framework can address this problem by introducing soft labels encoding the prior topic distributions of documents. The key, therefore, is to generate the soft labels based on the content of each document. Following MinV, we can first build a *topic representation* similar to the *literal usage representation*. Then, the semantic distance between a document and topic representations can be used to generate the soft labels. In addition, it is also worth inducing a prior topic distribution for each word. All these prior information can be integrated into the learning process following the idea of informed Gibbs sampling. We expect the weakly informed topic modeling can learn more precise representations of topics and alleviate the post-hoc labeling.

Improve the Semantic Compatibility Model The notion of semantic compatibility is significant to many NLP applications. We have analyzed the limitations of CBOW and adapt it to model the semantic compatibility between a sense and a context. Improving the performance of this part is a priority of future work.

First, it would be interesting to experiment with advanced context representations recently proposed in the literature, such as BERT [Devlin et al., 2018] and ELMo [Peters et al., 2018]. These representations can efficiently encode different types of syntactic and semantic information and have significantly outperformed the state of the art of several challenging NLP problems, e.g., sentiment analysis, question answering and textual entailment.

Second, it would also be interesting to investigate the negative sampling in the training of



(a) LDA[Blei et al., 2003]

(b) Weakly informed LDA

Figure 10: Graphical model of LDA and weakly informed LDA. s is the prior topic distribution.

semantic compatibility model. We find that the negative sampling strategy is somewhat naive; the current sampling algorithm selects a word as a negative word based solely on its frequency in the corpus. This could have a negative impact on the quality of the trained model. We conjecture that the reasons are twofold. First, it is highly likely that the generated training examples are not very challenging for the models to learn meaningful semantic compatibility patterns. Second, it might sample words which are compatible with the context (i.e., they are not negative). Thus, generating hard negative examples can help further improve the performance of the learned model. A related technique is *hard negative mining*, which is actively studied in the machine learning community [Shrivastava et al., 2016, Hinami and Satoh, 2018, Shi et al., 2018].

Transfer Our Models to Other Langauge A lot of language might lack manually crafted lexical resources, such as the MRC Psycholinguistic Database Machine Usable Dictionary (which provides abstractness measurement of words) used in our unsupervised model. Therefore, an interesting question is how can we transfer our models trained on English to other resource-constrained languages. One work on metaphor detection shows that model transferring can effectively leverage the knowledge learned from English to Spanish, Farsi, and Russian [Tsvetkov et al., 2014]. We think this idea can also apply to idioms and future research along this line is promising.

7.2.2 Open Research Questions

Infer the Figurative Meanings of Idioms The models we have proposed in this thesis can tell whether an idiom is used figuratively or literally. A more challenging task is to infer the figurative meanings of idioms without relying on manually crafted resources such as idiom dictionary.

One promising approach is through the use of semantic compatibility model we have proposed in Chapter 5. We can first collect a number of figurative instances of the target idiom and find words that are compatible with the contexts; these words can then be used to approximate the figurative meaning of the idiom. However, it is possible that the figurative interpretation cannot be fully expressed by a single word. A more generalized model should be able to generate interpretations of variable length. An example technique to achieve this goal is the Encoder-Decoder architecture [Sutskever et al., 2014], as shown in Fig. 11. The encoder aims to represent the context of idioms, while the decoder exploits recurrent neural networks to generate the inferred figurative meaning.

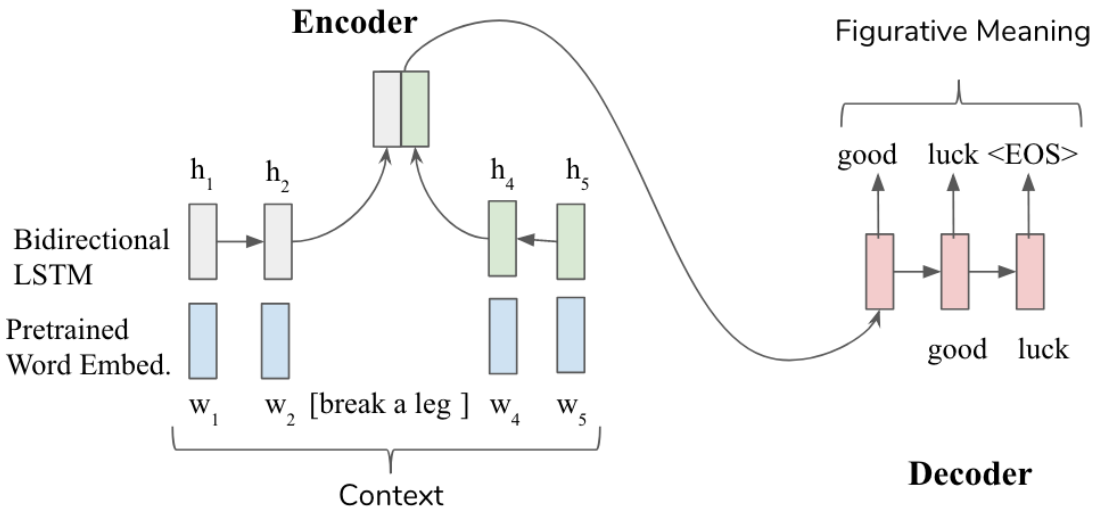


Figure 11: Encoder-Decoder model for inferring the figurative meanings of idioms.

To be able to interpret figurative languages is a longstanding problem in NLP. Inferring the figurative meaning of idioms opens up exciting research opportunities to address this challenging problem. We believe this is an important step toward seamless communication between human and computers. We will leave this as future work.

Idiom Recommendation Idioms is a major language barrier for non-native speakers. In a pilot study, we have surveyed seven non-native speakers on 100 Tweets containing idioms; we have found that, on average, the participants had trouble understanding 70% of them due to the inclusion of idioms. Communicating using idiom is also significant. Idioms often involve some cultural background knowledge thus they can convey certain subtle meaning in a concise and vivid way; non-native speakers who are not aware of the idioms might end up using plain and redundant language to describe the meaning which would otherwise be easily expressed by the idioms. Therefore, recommending idioms is useful for non-native speakers.

There are at least two types of idiom recommendations that are worth exploring. The first type is recommending an idiom purely based on meaning. This is useful when the users have an intended meaning they want to convey but they do not know what idioms to use. A related work is presented in [Hill et al., 2016]; the proposed model can recommend a word based on the sentences describing the meaning. The second type is recommending an idiom based on contexts. For example, when a user is writing an essay, it is of great value to build an intelligent idiom recommendation model that can locate parts of the writing which can be replaced by certain idioms. In this case, the contexts of the parts to be replaced provide useful information for the recommendation model.

To conclude this thesis, the research reported here demonstrates that linguistic-informed computational models capturing the interactions between idioms and contexts can help build robust and efficient idiom usage recognizers. Our model could benefit downstream NLP applications to alleviate the negative impact caused by the ambiguities of idiomatic expressions.

Bibliography

[A Bobrow and M Bell, 1973] A Bobrow, S. and M Bell, S. (1973). On catching on to idiomatic expressions. *Memory cognition*, 1:343–346.

[A. Swinney and Cutler, 1979] A. Swinney, D. and Cutler, A. (1979). The access and processing of idiomatic expressions. *Journal of Verbal Learning and Verbal Behavior*, 18:523–534.

[Abdalgader and Skabar, 2012] Abdalgader, K. and Skabar, A. (2012). Unsupervised similarity-based word sense disambiguation using context vectors and sentential word importance. *ACM Transactions on Speech and Language Processing (TSLP)*, 9(1):2.

[Agirre et al., 2006] Agirre, E., Martínez, D., de Lacalle, O. L., and Soroa, A. (2006). Two graph-based algorithms for state-of-the-art wsd. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 585–593. Association for Computational Linguistics.

[Ando, 2006] Ando, R. K. (2006). Applying alternating structure optimization to word sense disambiguation. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*, pages 77–84. Association for Computational Linguistics.

[Bahdanau et al., 2014] Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

[Balahur et al., 2013] Balahur, A., Steinberger, R., Kabadjov, M., Zavarella, V., Van Der Goot, E., Halkia, M., Pouliquen, B., and Belyaeva, J. (2013). Sentiment analysis in the news. *arXiv preprint arXiv:1309.6202*.

[Bannard, 2007] Bannard, C. (2007). A measure of syntactic flexibility for automatically identifying multiword expressions in corpora. In *Proceedings of the Workshop on a Broader Perspective on Multiword Expressions*, pages 1–8. Association for Computational Linguistics.

[Baroni et al., 2009] Baroni, M., Bernardini, S., Ferraresi, A., and Zanchetta, E. (2009). The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, 43(3):209–226.

[Bengio et al., 2013] Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828.

[Bengio et al., 2003] Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. (2003). A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155.

[Birke and Sarkar, 2006] Birke, J. and Sarkar, A. (2006). A clustering approach for nearly unsupervised recognition of nonliteral language. In *EACL*.

[Blei et al., 2003] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet Allocation. *The Journal of Machine Learning Research*, 3:993–1022.

[Bojar et al., 2017] Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huang, S., Huck, M., Koehn, P., Liu, Q., Logacheva, V., et al. (2017). Findings of the 2017 conference on machine translation (wmt17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214.

[Bruni et al., 2014] Bruni, E., Tran, N. K., and Baroni, M. (2014). Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, 49(1):1–47.

[Burnard, 2007] Burnard, L. (2007). Reference guide for the british national corpus. http://www.natcorp.ox.ac.uk/docs/URG/.

[Byrne et al., 2013] Byrne, L., Fenlon, C., and Dunnion, J. (2013). IIRG: A naive approach to evaluating phrasal semantics. *In Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation*, 45(4).

[Cacciari and Levorato, 1998] Cacciari, C. and Levorato, M. C. (1998). The effect of semantic analyzability of idioms in metalinguistic tasks. *Metaphor and Symbol*, 13(3):159–177.

[Cacciari et al., 1993] Cacciari, C., Tabossi, P., et al. (1993). Idioms. processing, structure and interpretation.

[Cap et al., 2015] Cap, F., Nirmal, M., Weller, M., and Im Walde, S. S. (2015). How to account for idiomatic german support verb constructions in statistical machine translation. In *Proceedings of the 11th Workshop on Multiword Expressions*, pages 19–28.

[Chen and Manning, 2014] Chen, D. and Manning, C. (2014). A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 740–750.

[Cheng et al., 2016] Cheng, J., Dong, L., and Lapata, M. (2016). Long short-term memory-networks for machine reading. *arXiv preprint arXiv:1601.06733*.

[Church et al., 1991] Church, K., Gale, W., and Hanks, P. (1991). Using statistics in lexical analysis. *Lexical acquisition: exploiting on-line resources to build a lexicon*, 115:164.

[Church and Hanks, 1990] Church, K. W. and Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29.

[Cilibrasi and Vitanyi, 2007] Cilibrasi, R. L. and Vitanyi, P. (2007). The google similarity distance. *Knowledge and Data Engineering, IEEE Transactions on*, 19(3):370–383.

[Collins, 2002] Collins, M. (2002). Discriminative training methods for Hidden Markov Models: Theory and experiments with perceptron algorithms. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 1–8. Association for Computational Linguistics.

[Coltheart, 1981] Coltheart, M. (1981). The mrc psycholinguistic database. *The Quarterly Journal of Experimental Psychology*, 33(4):497–505.

[Cook et al., 2007] Cook, P., Fazly, A., and Stevenson, S. (2007). Pulling their weight: Exploiting syntactic forms for the automatic identification of idiomatic expressions in context. In *Proceedings of the workshop on a broader perspective on multiword expressions*, pages 41–48. Association for Computational Linguistics.

[Cook et al., 2008] Cook, P., Fazly, A., and Stevenson, S. (2008). The vnc-tokens dataset. *Proc. of MWE (2008)*, pages 19–22.

[Cowie et al., 1984] Cowie, A. P., Mackin, R., and McCaig, I. R. (1984). Oxford dictionary of current idiomatic english, vol. i-ii. general introduction. Oxford University Press.

[David et al., 2005] David, G., Junbo, K., Ke, C., and Kazuaki, M. (2005). English gigaword second edition ldc2005t12. *Linguistic Data Consortium*.

[Devlin et al., 2018] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[Di Marco and Navigli, 2013] Di Marco, A. and Navigli, R. (2013). Clustering and diversifying web search results with graph-based word sense induction. *Computational Linguistics*, 39(3):709–754.

[F Bowdle and Gentner, 2005] F Bowdle, B. and Gentner, D. (2005). The career of metaphor. *Psychological review*, 112:193–216.

[Fadaee et al., 2018] Fadaee, M., Bisazza, A., and Monz, C. (2018). Examining the tip of the iceberg: A data set for idiom translation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.

[Fan et al., 2008] Fan, R. E., Chang, K. W., Hsieh, C. J., Wang, X. R., and Lin, C. J. (2008). Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874.

[Fass, 1991] Fass, D. (1991). met*: A method for discriminating metonymy and metaphor by computer. *Computational Linguistics*, 17(1):49–90.

[Fazly et al., 2009] Fazly, A., Cook, P., and Stevenson, S. (2009). Unsupervised type and token identification of idiomatic expressions. *Computational Linguistics*, 35(1):61–103.

[Fazly and Stevenson, 2006] Fazly, A. and Stevenson, S. (2006). Automatically constructing a lexicon of verb phrase idiomatic combinations. In *EACL*.

[Fellbaum, 1998] Fellbaum, C. (1998). *WordNet: An electronic lexical database (Language, Speech, and Communication). Cambridge, MA: The MIT Press*.

[Finkel et al., 2005] Finkel, J. R., Grenager, T., and Manning, C. (2005). Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, pages 363–370. Association for Computational Linguistics.

[Gagliano et al., 2016] Gagliano, A., Paul, E., Booten, K., and Hearst, M. A. (2016). Intersecting word vectors to take figurative language to new heights. In *Proceedings of the Fifth Workshop on Computational Linguistics for Literature*, pages 20–31.

[Gentner and Wolff, 1997] Gentner, D. and Wolff, P. (1997). Alignment in the processing of metaphor. *Journal of Memory and Language - J MEM LANG*, 37:331–355.

[Gibbs et al., 1989] Gibbs, R. W., Nayak, N. P., and Cutting, C. (1989). How to kick the bucket and not decompose: Analyzability and idiom processing. *Journal of memory and language*, 28(5):576–593.

[Gibbs Jr, 1992] Gibbs Jr, R. W. (1992). What do idioms really mean? *Journal of memory and language*, 31(4):485.

[Goatly, 1997] Goatly, A. (1997). *The language of metaphors*. Routledge.

[Graves et al., 2013] Graves, A., Jaitly, N., and Mohamed, A.-r. (2013). Hybrid speech recognition with deep bidirectional lstm. In *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*, pages 273–278. IEEE.

[Graves and Schmidhuber, 2005] Graves, A. and Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, 18(5):602–610.

[Griffiths and Steyvers, 2004] Griffiths, T. L. and Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National academy of Sciences*, 101(suppl 1):5228–5235.

[Halliday and Hasan, 2014] Halliday, M. A. K. and Hasan, R. (2014). *Cohesion in English*. Routledge.

[Harris, 1954] Harris, Z. S. (1954). Distributional structure. *Word*, 10(2-3):146–162.

[Hill et al., 2016] Hill, F., Cho, K., Korhonen, A., and Bengio, Y. (2016). Learning to understand phrases by embedding the dictionary. *Transactions of the Association for Computational Linguistics*, 4:17–30.

[Hinami and Satoh, 2018] Hinami, R. and Satoh, S. (2018). Discriminative learning of open-vocabulary object retrieval and localization by negative phrase augmentation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2605–2615, Brussels, Belgium. Association for Computational Linguistics.

[Hochreiter and Schmidhuber, 1997] Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

[Hristea, 2013] Hristea, F. T. (2013). *The Naïve Bayes Model in the Context of Word Sense Disambiguation*, pages 9–16. Springer Berlin Heidelberg, Berlin, Heidelberg.

[Jackendoff, 1997] Jackendoff, R. (1997). *The architecture of the language faculty*. Number 28. MIT Press.

[Kågebäck and Salomonsson, 2016] Kågebäck, M. and Salomonsson, H. (2016). Word sense disambiguation using a bidirectional lstm. *arXiv preprint arXiv:1606.03568*.

[Karov and Edelman, 1998] Karov, Y. and Edelman, S. (1998). Similarity-based word sense disambiguation. *Computational Linguistics*, 24(1):41–59.

[Katz and Giesbrecht, 2006] Katz, G. and Giesbrecht, E. (2006). Automatic identification of non-compositional multi-word expressions using latent semantic analysis. In *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, pages 12–19. Association for Computational Linguistics.

[Kingma and Ba, 2015] Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. *ICLR*.

[Kintsch, 2000] Kintsch, W. (2000). Metaphor comprehension: A computational theory. *Psychonomic Bulletin Review*, 7(2):257.

[Klein et al.,] Klein, G., Kim, Y., Deng, Y., Senellart, J., and Rush, A. M. OpenNMT: Open-Source Toolkit for Neural Machine Translation. *ArXiv e-prints*.

[Korkontzelos et al., 2013] Korkontzelos, I., Zesch, T., Zanzotto, F. M., and Biemann, C. (2013). Semeval-2013 task 5: Evaluating phrasal semantics.

[L. Hamblin and Gibbs, 1999] L. Hamblin, J. and Gibbs, R. (1999). Why you can't kick the bucket as you slowly die: Verbs in idiom comprehension. *Journal of Psycholinguistic Research*, 28:25–39.

[Lakoff, 1990] Lakoff, G. (1990). The invariance hypothesis: Is abstarct reason based on image-schemas? *Cognitive Linguistics*.

[Lakoff and Johnson, 1980] Lakoff, G. and Johnson, M. (1980). Metaphors we live by. *Chicago, IL: University of Chicago*.

[Landauer and Dutnais, 1997] Landauer, T. K. and Dutnais, S. T. (1997). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *PSYCHOLOGICAL REVIEW*, 104(2):211–240.

[Le and Mikolov, 2014] Le, Q. V. and Mikolov, T. (2014). Distributed representations of sentences and documents. *arXiv preprint arXiv:1405.4053*.

[Levy et al., 2015] Levy, O., Goldberg, Y., and Dagan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

[Li et al., 2010] Li, L., Roth, B., and Sporleder, C. (2010). Topic models for word sense disambiguation and token-based idiom detection. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1138–1147. Association for Computational Linguistics.

[Li and Sporleder, 2009] Li, L. and Sporleder, C. (2009). Classifier combination for contextual idiom detection without labelled data. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pages 315–323. Association for Computational Linguistics.

[Li and Sporleder, 2010] Li, L. and Sporleder, C. (2010). Linguistic cues for distinguishing literal and non-literal usages. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 683–691. Association for Computational Linguistics.

[Li et al., 2016] Li, P., Li, W., He, Z., Wang, X., Cao, Y., Zhou, J., and Xu, W. (2016). Dataset and neural recurrent sequence labeling model for open-domain factoid question answering. *arXiv preprint arXiv:1607.06275*.

[Lin, 1999] Lin, D. (1999). Automatic identification of non-compositional phrases. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 317–324. Association for Computational Linguistics.

[Lin et al., 2017] Lin, Z., Feng, M., Santos, C. N. d., Yu, M., Xiang, B., Zhou, B., and Bengio, Y. (2017). A structured self-attentive sentence embedding. *ICLR*.

[Liu and Hwa, 2016] Liu, C. and Hwa, R. (2016). Phrasal substitution of idiomatic expressions. In *HLT-NAACL*, pages 363–373.

[Liu and Hwa, 2017] Liu, C. and Hwa, R. (2017). Representations of context in recognizing the figurative and literal usages of idioms.

[Liu and Hwa, 2018] Liu, C. and Hwa, R. (2018). Heuristically informed unsupervised idiom usage recognition. In *Proceedings of Empirical Methods in Natural Language Processing*.

[Liu and Hwa, 2019] Liu, C. and Hwa, R. (2019). A generalized idiom usage recognition model based on semantic compatibility. In *Proceedings of The 33st AAAI Conference on Artificial Intelligence*.

[Liu et al., 2017] Liu, P., Qian, K., Qiu, X., and Huang, X. (2017). Idiom-aware compositional distributed semantics. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1204–1213.

[Lu et al., 2016] Lu, J., Yang, J., Batra, D., and Parikh, D. (2016). Hierarchical question-image co-attention for visual question answering. In *NIPS*, pages 289–297.

[Mason, 2004] Mason, Z. J. (2004). Cormet: a computational, corpus-based conventional metaphor extraction system. *Computational linguistics*, 30(1):23–44.

[Melamud et al., 2016] Melamud, O., Goldberger, J., and Dagan, I. (2016). context2vec: Learning generic context embedding with bidirectional lstm. In *CoNLL*, pages 51–61.

[Mihalcea and Faruque, 2004] Mihalcea, R. and Faruque, E. (2004). Senselearner: Minimally supervised word sense disambiguation for all words in open text. In *Proceedings of ACL/SIGLEX Senseval*, volume 3, pages 155–158.

[Mikolov et al., 2013a] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *Proceedings of Workshop at ICLR*.

[Mikolov et al., 2013b] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

[Mohammad and Hirst, 2006] Mohammad, S. and Hirst, G. (2006). Determining word sense dominance using a thesaurus. In *EACL*.

[Moon, 1998] Moon, R. (1998). Fixed expressions and idioms in english.

[Nigam et al., 2000] Nigam, K., McCallum, A. K., Thrun, S., and Mitchell, T. (2000). Text classification from labeled and unlabeled documents using em. *Machine learning*, 39(2):103–134.

[Nunberg et al., 1994] Nunberg, G., Sag, I. A., and Wasow, T. (1994). Idioms. *Language*, pages 491–538.

[Ong et al., 2014] Ong, N., Litman, D., and Brusilovsky, A. (2014). Ontology-based argument mining and automatic essay scoring. In *Proceedings of the First Workshop on Argumentation Mining*, pages 24–28.

[Pantel and Lin, 2002] Pantel, P. and Lin, D. (2002). Discovering word senses from text. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

[Papineni et al., 2002] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

[Pasini and Navigli, 2017] Pasini, T. and Navigli, R. (2017). Train-o-matic: Large-scale supervised word sense disambiguation in multiple languages without manual training data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 78–88.

[Patwardhan and Pedersen, 2006] Patwardhan, S. and Pedersen, T. (2006). Using wordnet-based context vectors to estimate the semantic relatedness of concepts. In *Proceedings of the eacl 2006 workshop making sense of sense-bringing computational linguistics and psycholinguistics together*, volume 1501, pages 1–8. Trento.

[Peng et al., 2014] Peng, J., Feldman, A., and Vylomova, E. (2014). Classifying idiomatic and literal expressions using topic models and intensity of emotions. *EMNLP*, pages 2019–2027.

[Persing and Ng, 2015] Persing, I. and Ng, V. (2015). Modeling argument strength in student essays. 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 543–552.

[Peters et al., 2018] Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, volume 1, pages 2227–2237.

[Pur and Pedersen, 2004] Pur, A. and Pedersen, T. (2004). Word sense discrimination by clustering contexts in vector and similarity spaces. *Proceedings of CoNLL-2004*.

[Rajani et al., 2014] Rajani, N. F., Salinas, E., and Mooney, R. (2014). Using abstract context to detect figurative language.

[Řehůřek and Sojka, 2010] Řehůřek, R. and Sojka, P. (2010). 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. ELRA. http://is.muni.cz/publication/884893/en.

[Rush et al., 2015] Rush, A. M., Chopra, S., and Weston, J. (2015). A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389.

[Salton et al., 2014] Salton, G., Ross, R., and Kelleher, J. (2014). An empirical study of the impact of idioms on phrase based statistical machine translation of english to brazilian-portuguese.

[Santos et al., 2016] Santos, C. d., Tan, M., Xiang, B., and Zhou, B. (2016). Attentive pooling networks. *arXiv preprint arXiv:1602.03609*.

[Schapire and Singer, 1999] Schapire, R. E. and Singer, Y. (1999). Improved boosting algorithms using confidence-rated predictions. *Machine Learning*, 37(3):297–336.

[Sennrich et al., 2016] Sennrich, R., Haddow, B., and Birch, A. (2016). Edinburgh neural machine translation systems for wmt 16. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, volume 2, pages 371–376.

[Shi et al., 2018] Shi, H., Mao, J., Xiao, T., Jiang, Y., and Sun, J. (2018). Learning visually-grounded semantics from contrastive adversarial samples. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3715–3727.

[Shrivastava et al., 2016] Shrivastava, A., Gupta, A., and Girshick, R. (2016). Training region-based object detectors with online hard example mining. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 761–769.

[Shutova, 2010a] Shutova, E. (2010a). Automatic metaphor interpretation as a paraphrasing task. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 1029–1037. Association for Computational Linguistics.

[Shutova, 2010b] Shutova, E. (2010b). Models of metaphor in nlp. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 688–697. Association for Computational Linguistics.

[Spasic et al., 2017] Spasic, I., Williams, L., and Buerki, A. (2017). Idiom—based features in sentiment analysis: Cutting the gordian knot. *IEEE Transactions on Affective Computing*.

[Sporleder and Li, 2009] Sporleder, C. and Li, L. (2009). Unsupervised recognition of literal and non-literal use of idiomatic expressions. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 754–762. Association for Computational Linguistics.

[Stone and Ann, 2016] Stone, S. and Ann, M. (2016). The difference between bucket-kicking and kicking the bucket: Understanding idiom flexibility.

[Sutskever et al., 2014] Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

[Tapanainen et al., 1998] Tapanainen, P., Piitulainen, J., and Järvinen, T. (1998). Idiomatic object usage and support verbs. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 2*, pages 1289–1293. Association for Computational Linguistics.

[Tsvetkov et al., 2014] Tsvetkov, Y., Boytsov, L., Gershman, A., Nyberg, E., and Dyer, C. (2014). Metaphor detection with cross-lingual model transfer. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 248–258.

[Turney et al., 2011] Turney, P. D., Neuman, Y., Assaf, D., and Cohen, Y. (2011). Literal and metaphorical sense identification through concrete and abstract context. In *Proceedings of*

the Conference on Empirical Methods in Natural Language Processing, pages 680–690. Association for Computational Linguistics.

[Turney and Pantel, 2010] Turney, P. D. and Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37:141–188.

[Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *arXiv preprint arXiv:1706.03762*.

[Veale and Hao, 2008] Veale, T. and Hao, Y. (2008). A fluid knowledge representation for understanding and generating creative metaphors. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 945–952. Association for Computational Linguistics.

[Venkatapathy and Joshi, 2005] Venkatapathy, S. and Joshi, A. K. (2005). Measuring the relative compositionality of verb-noun (vn) collocations by integrating features. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 899–906. Association for Computational Linguistics.

[Véronis, 2004] Véronis, J. (2004). Hyperlex: Lexical cartography for information retrieval. *Computer Speech Language*, 18:223–252.

[Williams et al., 2015] Williams, L., Bannister, C., Arribas-Ayllon, M., Preece, A., and Spasić, I. (2015). The role of idioms in sentiment analysis. *Expert Systems with Applications*, 42(21):7375–7385.

[Xiong et al., 2016] Xiong, C., Zhong, V., and Socher, R. (2016). Dynamic coattention networks for question answering. *arXiv preprint arXiv:1611.01604*.

[Xu et al., 2015] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., and Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057.

[Zhong and Ng, 2010] Zhong, Z. and Ng, H. T. (2010). It makes sense: A wide-coverage word sense disambiguation system for free text. In *Proceedings of the ACL 2010 System Demonstrations*, pages 78–83. Association for Computational Linguistics.

[Zweig and Burges, 2011] Zweig, G. and Burges, C. J. (2011). The microsoft research sentence completion challenge.