Master's Thesis

Unsupervised Word Sense Disambiguation Based on Word Embedding and Collocation

Han Shangzhuang

Supervisor Kiyoaki Shirai

Graduate School of Advanced Science and Technology
Japan Advanced Institute of Science and Technology
(Information Science)

March, 2021

Abstract

Word Sense Disambiguation (WSD) is a long-standing problem in Natural Language Processing, which aims to determine a sense for a target word in a given context. WSD plays a important role in downstream tasks of NLP, such as Machine Translation, Named Entity Recognition (NER), and chatbot. Approaches for WSD can be grouped into two main categories: methods based on supervised machine learning (supervised methods) and knowledge-based methods (unsupervised methods). Supervised methods often train a classifier or neural network model from a sense tagged corpus. Although supervised methods tend to achieve good performance, sense tagged corpora are required for training. Obviously, they are hard to construct due to heavy manual annotation. Knowledge-based WSD methods rely on lexical resources rather than sense tagged corpora. A gloss, which defines a meaning of a word in a dictionary, is first utilized in Lesk algorithm. Given a word and its context, Lesk algorithm calculates a score of each sense by measuring the number of overlapped words in a gloss (definition) of a sense and in a context. Then, the sense with the highest score is chosen. A lot of studies follow it and propose its extended models. One of the advantages of knowledge-based WSD methods is that it can be developed with existing lexical knowledge or database. Since a sense tagged corpus is not required, it can be easily implemented with low costs. However, the performance of the WSD tends to be lower than supervised approaches.

Word embedding is abstract representation of words in a form of an n-dimensional vector. It can be pre-trained from a large amount of corpus, then can be used for solving various NLP tasks. With the development of word embedding, some researchers propose other methods to compute the text similarity using word embedding instead of calculating overlaps in Lesk algorithm. For example, Basile et al. propose an unsupervised WSD algorithm which extends the Lesk's WSD method. In Basile's method, the word similarity in the semantic space is regarded as gloss-context overlap. Three steps are required to determine a sense of a target word: (1) to construct a context vector $\vec{c}$, (2) to construct sense vectors $\vec{s_i}$, and (3) to calculate the cosine similarity of two vectors to choose the sense. The context vector is obtained by averaging vectors of all context words in a context. Pre-trained word embedding is used as word vectors $\vec{w_k}$. Similarly, the sense vector is constructed by averaging the word vectors in a gloss sentence. Finally, the sense whose vector is the most similar to the context vector is chosen.

The goal of this thesis is to propose a novel unsupervised WSD method. We extend the Basile's method in two directions. One is to incorporate a mechanism to determine a sense using a collocation. Rules to determine a sense, which are based on collocations, are automatically acquired from a raw corpus, then these rules are integrated to the Basile's WSD model. Two types of the collocation are considered in this study. The first one is a word collocation that is a sequence of words including the target word. The second collocation is a dependency collocation, which is defined as a pair of words under a certain syntactic dependency. In this study, we call the rule "collocation WSD rule". The other extension is that we investigate the better way to construct the context vector in the Basile's method. Our basic idea is to use not all words but only words that are highly related to the sense for the construction of the context vector. We defined an indicator called *RelevantScore*, which evaluates how a word is strongly related to a particular sense. In this study, *RelevantScore* is obtained by the cosine similarity of the word vector (word embedding) and the sense vector. Then, the new context vector is made by averaging word vectors of words with high *RelevantScore*. Hereafter, we call the WSD method considering *RelevantScore* the Highly Related Word Embedding method (HRWE).

In our method, we use the HRWE and some filtering methods to extract high-quality collocation WSD rules from a raw corpus in advance, and store them in a database. For an ambiguous word in an unlabeled sentence in a row corpus, the sense of it is determined by the HRWE. If the reliability of the chosen sense is high, candidates of the collocation WSD rules are obtained by applying the pre-defined templates. Finally, rules are filtered out if they consist of only function words, do not frequently occur, or are inaccurate for WSD.

In our proposed system, when performing WSD, the sense of a target word is determined by the collocation WSD rule if the collocation in the rule is found in the context of the target word. Otherwise, the HRWE is used to determine the sense of a target word. The several ideas to improve the WSD performance are newly introduced in it: to construct the context vector with only contextual words that are highly related to the sense (HRWE), to acquire and combine the collocation WSD rules, to use two types of collocation WSD rules (word collocation and dependency collocation), and so on. Several experiments are conducted to discuss the contribution of each component in our method through the comparison of different WSD models.

The dataset of Senseval-3 English lexical sample task is used in the experiment. In addition to the test dataset, we also need two external data, one is a lexical resource as a sense inventory, the other is a raw corpus for extraction of the collocation WSD rules. We use WordNet and Leipzig corpus

as our sense inventory and raw corpus respectively.

Comparing the precision of WSD on the dataset, the HRWE outperforms the baseline for nouns and verbs, but not for adjectives. However, the precision is improved by 3.2 point for all POSs. It indicates that our idea to select contextual words strongly associated with senses for construction of the context vector is effective. The system using only the collocation WSD rules tends to achieve the higher precision than other systems with low applicability, especially for nouns and adjectives. It is confirmed that we can obtain the disambiguation rules whose recall is low but precision is high as we aimed. The performance of the systems integrating the HRWE with the word or dependency collocation WSD rules is better than the HRWE only, which means the collocations can contribute to choose the appropriate sense. Our final system, the WSD system with the HRWE and both the word and dependency collocation WSD rules, achieves the best performance for nouns, verbs and all POSs. Its precision is 0.572, which is 4.7 point better than the baseline.

Contents

1 Introduction **1**
- 1.1 Research Background . . . 1
- 1.2 Problem Statement . . . 2
- 1.3 Research Goal . . . 2
- 1.4 Organization of this Thesis . . . 3

2 Related Work **4**
- 2.1 Word Sense Disambiguation . . . 4
  - 2.1.1 Features for WSD . . . 5
  - 2.1.2 Supervised WSD . . . 5
  - 2.1.3 Unsupervised WSD . . . 7
- 2.2 Word Embedding . . . 8
- 2.3 Characteristics of this Study . . . 11

3 Proposed Method **12**
- 3.1 Overview of Method . . . 12
- 3.2 Highly Related Word Embedding Method . . . 12
  - 3.2.1 Basile's Method . . . 13
  - 3.2.2 Our Extension . . . 15
- 3.3 Collocation Based WSD . . . 17
  - 3.3.1 Collocation WSD Rule . . . 17
  - 3.3.2 Construction of Collocation WSD Rule . . . 19
  - 3.3.3 Filtering Collocation WSD Rule . . . 21
  - 3.3.4 Summary . . . 22

4 Evaluation **24**
- 4.1 Experimental Setting . . . 24
  - 4.1.1 WordNet . . . 24
  - 4.1.2 Data . . . 25
  - 4.1.3 Evaluation criteria . . . 28
- 4.2 Preliminary experiment . . . 29

4.2.1 Comparison of word embedding models 29
4.2.2 Evaluation of gloss expansion 30
4.3 Results 31
4.4 Evaluation of collocation WSD rules 33
4.5 Discussion about context window size 34

5 Conclusion **37**
5.1 Concluding Remark 37
5.2 Future Work 37

A Sense Inventory **42**

List of Figures

2.1 Example of WSD by GlossBERT [8] 6
2.2 Web-based variant of the Lesk algorithm [5] 7
2.3 Excerpt of the WordNet graph where $drink_v^1$ is centered [17] . 9
2.4 Skip-gram model [13] 10
2.5 The input of BERT 11

3.1 Overview of proposed WSD system 13
3.2 Example of context vector and sense vector 14
3.3 Relationship between synsets in WordNet 15
3.4 Example of relevant word set 17
3.5 Template of word collocation WSD rule 18
3.6 Template of dependency collocation WSD rule 18
3.7 Example of usage of collocation WSD rule 19
3.8 Flowchart of acquisition of collocation WSD rule. 19
3.9 Obtained rules from example sentence 20
3.10 Dependency tree of example sentence 21
3.11 Function word templates for filtering 22

4.1 Example of Search on WordNet 25
4.2 An example of test instance 26
4.3 Example of acquired collocation WSD rule 34
4.4 Precision of models with different context window sizes 36

List of Tables

4.1 Dataset of Senseval-3 English lexical sample task. 26
4.2 List of target words 26
4.3 Example of sense unification 27
4.4 Average number of senses 28
4.6 Comparison of word embedding. 30
4.5 Probing task performance for each BERT layer [9] 30
4.7 Evaluation of gloss expansion on Senseval-3 31
4.8 Comparison of Baseline and HRWE method 31
4.9 Results of the system using the collocation WSD rule only . . 32
4.10 Comparison of systems with or without collocation WSD rule 32
4.11 Comparison of WSD methods 33
4.12 Parameters for acquisition of collocation WSD rule 33
4.13 Number of rules mined from raw corpus 34

A.1 Original and unified senses of verb (1) 42
A.2 Original and unified senses of verb (2) 43
A.3 Original and unified senses of verb (3) 44
A.4 Original and unified senses of verb (4) 45
A.5 Original and unified senses of verb (5) 46
A.6 Original and unified senses of verb (6) 47
A.7 Original and unified senses of noun (1) 48
A.8 Original and unified senses of noun (2) 49
A.9 Original and unified senses of noun (3) 50
A.10 Original and unified senses of noun (4) 51
A.11 Original and unified senses of noun (5) 52
A.12 Original and unified senses of adjective (1) 52
A.13 Original and unified senses of adjective (2) 53
A.14 Original and unified senses of adjective (3) 54

Chapter 1

Introduction

1.1 Research Background

Word Sense Disambiguation (WSD) is a fundamental task and long-standing challenge in Natural Language Processing (NLP), which aims to determine a sense of an ambiguous word in a particular context [16]. WSD plays a important role in downstream tasks of NLP, such as Machine Translation, Named Entity Recognition (NER), and chat-bot. The WSD approaches can be grouped into two main categories: methods based on supervised machine learning (supervised methods) and knowledge-based methods (unsupervised methods).

Supervised methods often train a classifier or neural network model from a sense tagged corpus. SemCor corpus [15] is one of the sense tagged corpora often used in the research area of WSD. In the early days, Support Vector Machine (SVM) was often used for WSD. In recent years, Bidirectional Encoder Representations from Transformers (BERT) [4] has often achieved the state-of-the-art performance in many NLP tasks. As we will discuss in Subsection 2.2, BERT was also applied for WSD. Although these supervised methods tend to achieve good performance, sense tagged corpora are required for training. Obviously, they are hard to construct due to heavy manual annotation.

Knowledge-based WSD methods rely on lexical resources like a dictionary or WordNet [14] rather than sense tagged corpora. A gloss, which defines a meaning of a word in a dictionary, is first utilized in Lesk algorithm [11]. Given a word and its context, Lesk algorithm calculates a score of each sense by measuring the number of overlapped words in a gloss (definition) of a sense of a target word and that of words in a context. Then, the sense with the highest score is chosen. A lot of studies follow it and propose its ex-

tended models. In addition to methods using gloss sentences, a graph-based WSD method is also investigated. In this approach, graph nodes correspond to word senses, whereas edges represent dependencies between senses (e.g. synonymy and antonymy). Sense disambiguation process is done by finding the most "important" node in the graph. One of the advantages of knowledge-based WSD methods is that it can be developed with existing lexical knowledge or database. Since a sense tagged corpus is not required, it can be easily implemented with low costs. However, the performance of the WSD tends to be lower than supervised approaches.

Although the supervised and knowledge-based WSD have different characteristics, both are worth to be investigated. This thesis focuses on the knowledge-based WSD using a dictionary as lexical knowledge.

1.2 Problem Statement

In the dictionary-based WSD methods based on Lesk algorithm and its variants, we think there are two major problems.

First, it only relies on words in a context of a target word. However, it is well-known that a collocation is another useful feature for WSD. Collocation is a series of words or terms that frequently co-occur more than expected by chance. Words in a collocation usually have a special and fixed meaning. For example, the collocation "hot spring" indicates that the sense of "spring" is FOUNTAIN, not SEASON. Therefore, it is a major problem that the collocation is not considered in the dictionary-based WSD.

Second, there is much room to improve the way how to measure the similarity between a gloss and a context. In recent study, both a gloss and a context are represented by vectors, then similarity between two vectors, e.g. cosine similarity, is calculated. In the construction of the vector of the context, all words in the context are often taken into account. However, not all words are related to a sense of a target word. Some words may be not related to the sense and even give bad influence to WSD. Generally, in the algorithm based on a similarity between sentence pairs, high-quality vector representation is essential.

1.3 Research Goal

The objective of this research is to develop a stable and high-precision WSD system using unsupervised method based on word embedding and collocation feature. The word embedding is a vector of a word, which represents an

abstract meaning of the word. It is used to make a vector of a gloss and a context in this study. To achieve this goal, we set the following two sub-goals.

1. While past studies only consider words in a context for WSD, our method also takes collocations into account to determine a sense of a given word. In addition to the ordinary collocation (adjacent words that often appear together), we also define a dependency collocation, which is a syntactic dependency relation between a target word and another word in a sentence.

2. We also propose a better way how to make a context vector. Our method only considers words that are highly related to the sense when the context vector is built using the word embedding.

The effectiveness of collocation features and the new context vector construction method will be empirically evaluated via several experiments.

1.4 Organization of this Thesis

The rest of the thesis is organized as follows. Chapter 2 introduces related work and clarifies characteristics of this study. Chapter 3 describes the details of our proposed method, including a refined method to build a context vector and a procedure to extract high quality WSD rules using collocations from a raw corpus. Chapter 4 reports several experiments to evaluate our method. Finally, Chapter 5 concludes this thesis.

Chapter 2

Related Work

This chapter introduces previous studies that are related to our study. Section 2.1 presents the Word Sense Disambiguation task and related work for it. Section 2.2 introduces three kinds of word embeddings. Finally, Section 2.3 discusses the characteristics of our method.

2.1 Word Sense Disambiguation

In many natural languages, a meaning of a word is ambiguous. It means that words can be interpreted as different meanings (senses) depending on contexts. Here are examples:

(a) The **spring** was broken. → a metal elastic device
(b) **Spring** is the best season of the year. → the season of growth

Obviously, the word "spring" has different senses in these two sentences as indicated in right of the arrow. Humans can easily distinguish these differences, however, it is difficult for a computer to determine a correct sense. Recognizing a correct sense of a word in a context with computational method is called Word Sense Disambiguation (WSD). Actually, WSD is considered an AI-complete problem [12], which means computers can solve these problems when they are as smart as humans. Researchers have made a lot of attempts in this field. At present, there are mainly two types of methods to solve this task: one is based on supervised learning, the other is based on knowledge resources. Both types of methods rely on several high-quality features. Commonly used features for WSD are firstly introduced in Subsection 2.1.1. Then, the previous work of the supervised and knowledge-based WSD methods are introduced in Subsection 2.1.2 and 2.1.3, respectively.

2.1.1 Features for WSD

There are three commonly used features in WSD. The first one is words in the surroundings of the target word. It is a bag of disordered words, which can determine the topic of the context. Since words that appear in the similar context tend to have the same meaning, it is considered an effective feature. The second feature is Part-of-speech (POS). The POS tags of the neighboring words are widely used. POS of an ambiguous word itself is also an effective feature, since a word has different senses under different POSs. The third feature is local collocations, which represent another standard feature that captures the ordered sequence of words which tend to appear around the target word [2].

2.1.2 Supervised WSD

Training a machine learning classifier or neural network model is a common way to solve WSD problem. It is called supervised WSD, since it is based on the supervised machine learning using the labeled data (i.e. a sense tagged corpus in the case of WSD). Le and Shimazu propose a method based on Naive Bayes [10]. A context of an ambiguous word is represented as a feature vector $\vec{F} = (f_1, f_2, ..., f_n)^T$ and the senses of the ambiguous word is represented as $(s_1, s_2, ..., s_k)$. Choosing the correct sense is finding the sense s_i that maximizes the conditional probability $P(s_i|\vec{F})$ as shown in Equation (2.1).

$$s' = \arg\max_{s_i} P(s_i|\vec{F}) \tag{2.1}$$

SVM is another machine learning algorithm that is widely used in WSD. The SVM finds a hyperplane with the greatest margin separating the training samples into two classes in a space of feature vectors of the samples. The instances in the same side of the hyperplane have the same class label. The features of test instances decide which side of the hyperplane the instance is located. Although SVM is a binary classification algorithm, it can be extended to tackle multi-class problems. Guo et al. try to utilize SVM algorithm to WSD [6]. Their system consists of two parts, feature extraction and classification. In the first step, four features (surrounding words, POS, collocation and syntactic relation) are extracted from instances. In the second step, an SVM classifier is trained from a set of feature vectors extracted in the first step.

Neural network model can be regarded as another type of a classifier. It has extraordinary feature extraction capabilities, although a number of

parameters is quit huge. Correspondingly, the neural network model requires a lot of annotated data to ensure the quality of the model. A solution to this problem is a pre-trained language model. From a huge amounts of texts, characteristics of a language are learnt to a certain extent in advance. Then the pre-trained language model is applied to downstream tasks with a few labeled data. It can achieve good results even without a lot of labeled data.

BERT [4] is a popular language model in recent years because it has achieved the state-of-the-art performance in many NLP tasks, such as question answering and language inference. Since the training process of BERT is based on two unsupervised tasks, Masked Language Model and Next Sentence Prediction, BERT can identify the relation between two sentences. Based on this idea, Huang et al. propose the GlossBERT, which constructs context-gloss pairs from all possible senses of the target word in WordNet, then treats the WSD task as a sentence-pair classification problem [8]. Figure 2.1 shows how the GlossBERT works. In this example, the target word is

Sentence with four targets:
Your research stopped when a convenient assertion could be made.

Context-Gloss Pairs of the target word [research]	Label	Sense Key
[CLS] Your research ... [SEP] systematic investigation to ... [SEP]	Yes	research%1:04:00::
[CLS] Your research ... [SEP] a search for knowledge [SEP]	No	research%1:09:00::
[CLS] Your research ... [SEP] inquire into [SEP]	No	research%2:31:00::
[CLS] Your research ... [SEP] attempt to find out in a ... [SEP]	No	research%2:32:00::

Context-Gloss Pairs with weak supervision of the target word [research]	Label	Sense Key
[CLS] Your "research" ... [SEP] research: systematic investigation to ... [SEP]	Yes	research%1:04:00::
[CLS] Your "research" ... [SEP] research: a search for knowledge [SEP]	No	research%1:09:00::
[CLS] Your "research" ... [SEP] research: inquire into [SEP]	No	research%2:31:00::
[CLS] Your "research" ... [SEP] research: attempt to find out in a ... [SEP]	No	research%2:32:00::

Figure 2.1: Example of WSD by GlossBERT [8]

"research" and the context is "Your research stopped when ..." at the top in the table. For each target word, N possible sense glosses are extracted from WordNet. In this case, four senses of "research" are obtained. Then, a pair of the context sentence and the gloss sentence of each sense forms "Context-Gloss" pair with special tokens [CLS] and [SEP]. In BERT, [CLS] is used to get abstract representation of two sentences, while [SEP] is a separator of two sentences. Next, the label "Yes" is assigned to the correct sense and "No" to the other senses. This data is used as the training data for fine-tuning of the BERT. When testing, the system outputs the probability of $label = yes$ of each "Context-Gloss" pair and choose the sense with the highest probability as the correct label of the target word.

2.1.3 Unsupervised WSD

The unsupervised WSD can be used in practice because it does not require annotated data. Unsupervised WSD methods often rely on lexical resources like a dictionary or WordNet [14]. The former is based on textual similarity, while the latter is based on a graph.

Lesk proposes a method which can determine the sense of target word using only the context and the gloss [11]. As already introduced in Section 1.1, Lesk algorithm calculates a score of each sense by measuring the number of overlapped words in a gloss (definition) of a sense of a target word and that of words in a context. Then, the sense with the highest score is chosen. A lot of studies follow it and propose its extended models.

Gaona et al. propose a variant algorithm of the Lesk approach, which is based on the hypothesis that the gloss of a sense and the context are highly related [5]. It measures the relationship of them by calculating the frequencies of co-occurrence words between the gloss and the context using Web. Figure 2.2 shows the pseudo code of this method. They use a query to a search engine to get a hit number of a set of words, called "web frequency". "d" means a set of words in a gloss and example sentences of a sense, "c" means a set of words in a context, and "dc" means a set of words in either "d" or "c". The weight means the probability of seeing the gloss of a sense in the given context. The sense which maximizes the weight is regarded as the answer.

```
For each word w to be tagged
 For each sense s of w
  g = gloss of sense s (bag of words)
  e = example of sense s (bag of words)
  d = g ∪ e
  dc = d ∪ c
  f_g  = web frequency of d
  f_gc = web frequency of dc
  weight(s) = f_gc/f_g
s = argmax weight (s)
```

Figure 2.2: Web-based variant of the Lesk algorithm [5]

With the development of word embedding, some researchers began to use other methods to compute the text similarity instead of calculating overlaps of words. For example, Basile et al. propose an unsupervised WSD algorithm which extends the Lesk's WSD method [1]. The most important

contribution of this research is a distributional semantic space is introduced to Lesk algorithm. The word similarity in the semantic space is regarded as gloss-context overlap. The semantic space is geometrical space of words where vectors express concepts of words, and proximity in the space is used to measure semantic relatedness between two words or two sentences. The semantic space is formed by word embedding pre-trained from a large amount of texts. For a given target word and its context, a context vector and gloss vector are constructed, then the similarity between them is calculated to determine the correct sense of the target word in the context. The details of Basile's algorithm will be introduced in Subsection 2.2. Actually, the idea of semantic spaces is inspired by the distributional hypothesis that was put forward very early [7]. The hypothesis supposes that words are semantically similar if they share contexts (surrounding words).

In addition to text similarity based methods, a graph-based WSD method is also investigated [17]. It consists of two stages. First, building a graph from a lexical resource representing all possible interpretations (senses) of the target word. In the graph, nodes correspond to senses, whereas edges represent relationship between senses (e.g., synonymy and antonotmy). Second, the graph structure is assessed to calculate the significance (or importance) of each node. WSD is performed by finding the most "significant" node for each word. Figure 2.3 shows an excerpt of the WordNet graph where $drink_v^1$ is centered. In this graph, the node of $drink_v^1$, the first sense of the verb "drink", is shown in black with dark gray background, and its adjacent nodes (senses) are in black. Senses which are not directly connected to $drink_v^1$ but reachable through other edges are shown in light gray. Two senses are connected if a relation is defined in WordNet. The correct sense is selected by ranking each vertex in the graph according its significance.

2.2 Word Embedding

Word embedding is abstract representation of words in a form of an n-dimensional vector. It can be pre-trained form a large amount of corpus, then can be used for solving various NLP tasks. Since word embedding plays an important role in our proposed WSD method, the overview of it is introduced in this section. Skip-gram [13] and Glove [18] are two typical models of training word embeddings.

Figure 2.4 shows the structure of the Skip-gram. The model is a simple feed-forward neural network that contains one input layer, one hidden layer (projection) and one output layer. $w(t)$ is a target word and $w(t-2), w(t-1), w(t+1), w(t+2)$ are context words of the target word. t stands for

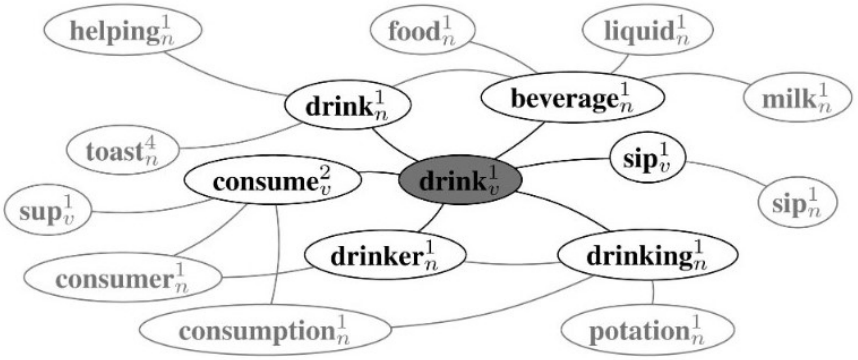


Figure 2.3: Excerpt of the WordNet graph where $drink^1_v$ is centered [17]

a position of a word. The objective of Skip-gram model is to maximize the probability of appearance of surrounding words when the target word is given. During training, an input word $w(t)$ is represented as a one-hot vector. This vector has N dimensions (N is the size of the vocabulary). A value of the vector is set to 1 at the position corresponding to the word $w(t)$, and 0 at all of the other positions. Hidden layer is a $N * M$ matrix (M is the specified word vector dimension). The output of the network is also a N-dimensional vector, that represents surrounding words. This neural network can be trained from a raw corpus that is a collection of a target word (input) and its surrounding words (output). After the training, word embedding (M-dimensional vector of a word) is obtained from the signals of the hidden layer.

We can see that this training procedure only takes into account the local information of the text. Glove overcomes this shortcoming. It leverages statistical information by training only on the non-zero elements in a word-word co-occurrence matrix, rather than on individual context windows in a large corpus.

Both Skip-gram and Glove can only generate a vector for a word, not a sense. That is, word embedding cannot distinguish different senses of a word. Only one vector is obtained for one word even when it has two or more senses.

Different from the previous two methods, language models trained based on Deep Neural Network Models can generate contextually dependent word vectors. BERT is one of the such language models. Its architecture is a multi-layer bidirectional Transformer encoder. Figure 2.5 shows an input of BERT. It is composed of three type of embeddings. “Token Embeddings” are

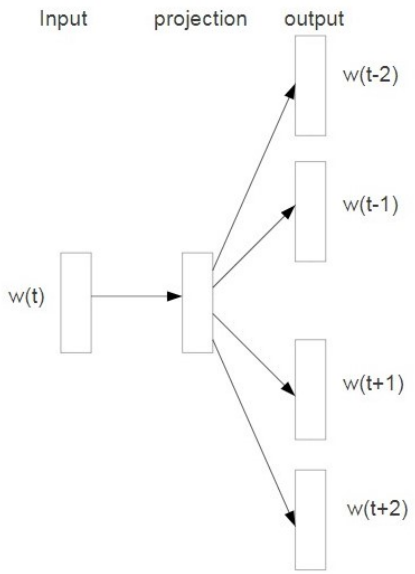


Figure 2.4: Skip-gram model [13]

the vector representations of words, which are similar to word embeddings. "Segment Embeddings" are vector representations to help BERT distinguish between paired input sequences. BERT can classify either a single sentence (e.g. polarity classification) or a pair of sentences (e.g. textual entailment). The segment embeddings works in the latter case. "Position Embeddings" lets BERT know that the inputs have a temporal property as the input is often regarded as a time sequence.

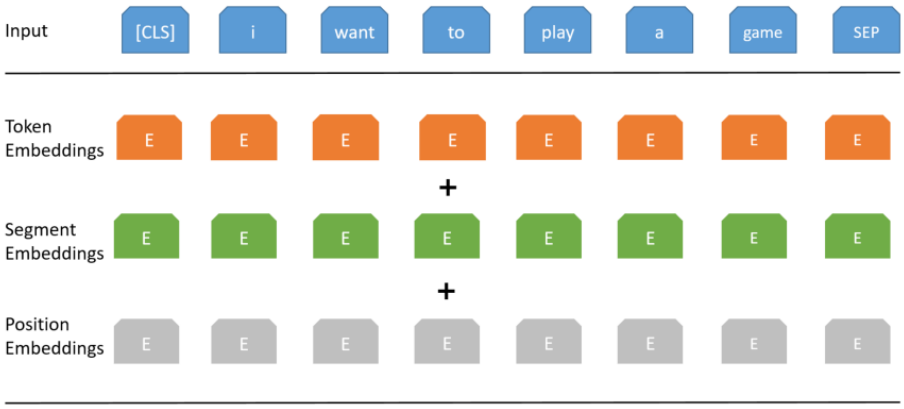


Figure 2.5: The input of BERT

The training of the BERT consists of two steps. The first step is pre-training of a language model. It is performed on a large corpus by conducting two novel unsupervised prediction tasks, the masked language model task and next sentence prediction task. The second step is fine-tuning of the model. The parameters of BERT are updated on a relatively small amount of annotated data for a downstream task. It is known that a pre-trained BERT can produce appropriate embedding of a sentence or a word. As Skip-gram and Glove, word embedding of the pre-trained BERT can also be used for WSD. Note that the fine-tuning requires a labeled data, but pre-training can be done using unlabeled data. Thus a system using pre-trained BERT is also an unsupervised WSD model.

2.3 Characteristics of this Study

Our method is unsupervised and text similarity based method, which is an improved version of Basile's method. This paper extends the Basile's method in two directions. One is to incorporate a mechanism to determine a sense using a collocation. Rules to determine a sense, which are based on collocations, are automatically acquired from a raw corpus, then these rules are integrated to the Basile's WSD model. The other is to propose a better way to construct the context vector, which can ignore noisy words in the context.

Chapter 3

Proposed Method

3.1 Overview of Method

Figure 3.1 shows an overview of the proposed system. It accepts a sentence including a target word as an input and chooses a sense of it as an output.

Our system consists of two modules: one is a rule based WSD system, the other is WSD system based on Highly Related Word Embedding. Hereafter, the latter is denoted as "HRWE method" in short. The first module uses the database of collocation WSD rules, which determine the sense by a collocation (word sequence). Briefly, these rules determine the sense by a collocation as *collocation* $\rightarrow$ *sense*. If a rule is hit for a collocation in a given sentence, the sense is chosen by the rule, otherwise the next module is applied. The second module is similar to the Basile's method [1]. It measures the similarity between gloss sentences in a dictionary and a context of a target word in a given sentence, then choose the sense whose gloss is the most similar to the context of the target word. Since the rule-based module is designed to achieve high precision in compensation for low recall, it is applied first.

In the following sections, the HRWE method will be introduced first, since it is also used to construct the sets of the collocation WSD rules. Then, the rule based WSD system is described, especially how to acquire WSD rules automatically.

3.2 Highly Related Word Embedding Method

Our algorithm is extended based on the Basile's method. In the next subsection, we introduce the Basile's method, and in the Subsection 3.2.2, we introduce the proposed method.

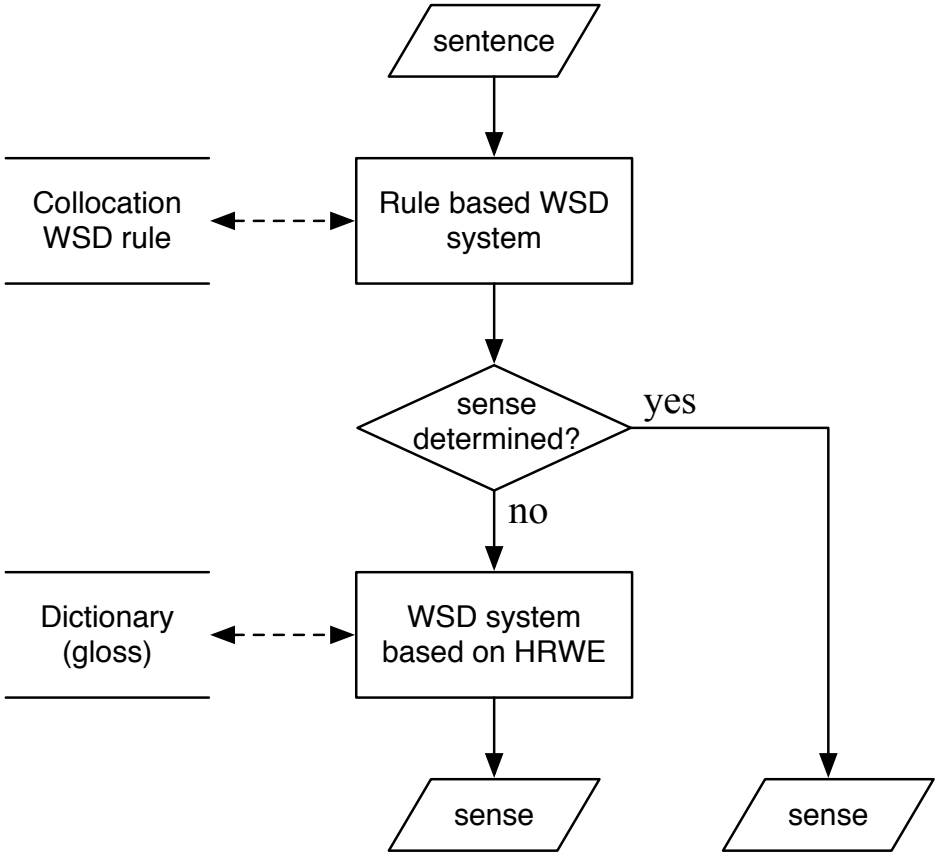


Figure 3.1: Overview of proposed WSD system

3.2.1 Basile's Method

In Basile's method [1], three steps are required to determine a sense of a target word: (1) to construct a context vector $\vec{c}$, (2) to construct sense vectors $\vec{s_i}$, and (3) to calculate the cosine similarity of two vectors to choose the sense. The context vector is obtained by averaging vectors of all context words in a context as Equation (3.1),

$$\vec{c} = \frac{1}{|W|} \sum_{w_k \in W} \vec{w_k} \tag{3.1}$$

where W stands for a set of words in the context. Pre-trained word embedding is used as word vectors $\vec{w_k}$. Similarly, the sense vector is constructed by averaging the word vectors in a gloss sentence as Equation (3.2).

$$\vec{s_i} = \frac{1}{|G_i|} \sum_{w_k \in G_i} \vec{w_k} \tag{3.2}$$

G_i is a set of words in a gloss sentence of the i-th sense.

Finally, as show in Equation (3.3), the sense whose vector is the most similar to the context vector is chosen.

$$s = \arg\max_{s_i} \ cos(\vec{c}, \vec{s_i}) \tag{3.3}$$

An important parameter in this method is the context window size, *CWS*. When constructing the context vector, the most nearest *CWS* words appearing before and after the target word are taken into account. Note that function words are ignored. That is, a content word is used to make a context vector if it is one of the most closest *CWS* content words, even when the distance between it and a target word is greater than *CWS*. On the other hand, *CWS* is not considered in the construction of the sense vector. That is, all words in a gloss is used. This is because that gloss sentences in WordNet, which is used as a dictionary in Basile's and our study, are rather short and concise.

Figure 3.2 shows examples of the context vector $\vec{c}$ and sense vector $\vec{s_i}$. Here we suppose the *CWS* is 3. In the context, the underlined word "bank" is the target word, and words in bold are the content words in the context window whose size is 3. Then, $\vec{c}$ is obtained by the average of word embeddings of these words. Similarly, the sense vector of the first sense $\vec{s}$ is the average of word embeddings of the words in the gloss sentence $Gloss_1$. In this case, all contents words are taken into account.

Figure 3.2: Example of context vector and sense vector

Context:	The **Consumer Federation claims** banks are **ripping** you off by not **passing along** savings on interest rates.
$\vec{c}$:	$\frac{1}{6} * (\overrightarrow{consumer} + \overrightarrow{federation} + \overrightarrow{claim} + \overrightarrow{rip} + \overrightarrow{pass} + \overrightarrow{along})$
Gloss_1:	an **arrangement** of **similar objects** in a **row**
$\vec{s_1}$:	$\frac{1}{4} * (\overrightarrow{arrangement} + \overrightarrow{similar} + \overrightarrow{objects} + \overrightarrow{row})$

As already explained, gloss sentences in WordNet are used in the Basile's study, but they are rather concise. To enrich sense vectors, Basile expanded the gloss using an API provided by BabelNet, which can extract all senses related to a particular sense. In this study, we expand the sense information with the gloss of the hypernyms, hyponyms and synonyms, and empirically evaluate its effectiveness in the experiment. Figure 3.3 shows an example of expansion of the gloss for a target word *bank* when it is a noun and its sense id is "bank.n.02". The sense definition in WordNet for "bank.n.02" is *a financial*

institution that accepts deposits and channels the money into lending activities. In addition to this sentence, the hypernyms (financial_institution.n.01) and hyponyms (acquirer.n.02, agent_bank.n.02, credit_union.n.01, and state_bank.n.01) are obtained by the BabelNet API, then gloss sentences of these senses are used to construct the sense vector of "bank.n.02".

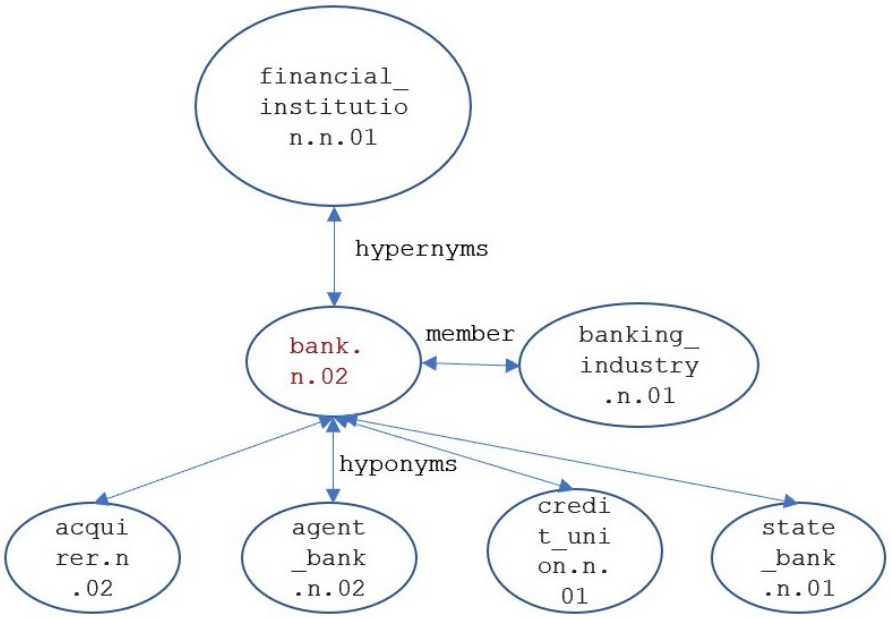


Figure 3.3: Relationship between synsets in WordNet

3.2.2 Our Extension

We investigate a better way to construct the context vector in Basile's method. Our basic idea is to use not all words but only words that are highly related to the sense for the construction of the context vector.

When constructing a text vector, averaging the word vector of each word under the unit window is a common method. An important issue is how to determine an appropriate window size, since there are pros and cons of a large or small window. When the window size is too small, it is likely that not enough information is obtained in the context vector. However, when the window size is too large, many unrelated words are added to the context vector. It may decrease the quality of it. Our Highly Related Word

Embedding method can ensure that words with close meanings to the target word are selected even in a large window size.

For each sense s_i, a different context vector, denoted as $\vec{c^{(i)}}$, is made from contextual words relevant to s_i. First, for each word w_k in a context, the relevance score in terms of the i-th sense is defined as Equation (3.4).

$$RelevantScore(w_k^{(i)}) = cos(\vec{w_k}, \vec{s_i}) \tag{3.4}$$

We assume that the word with high *RelevantScore* is strongly related to the particular sense, thus it is effective feature for WSD. The relevant word set, $WR^{(i)}$, is made by selecting the top T_r words with the highest *RelevantScore* for each sense s_i. Then the new sense-dependent context vector is made by averaging word vectors of words in $WR^{(i)}$ as Equation (3.5).

$$\vec{c^{(i)}} = \frac{1}{|WR^{(i)}|} \sum_{w_k \in WR^{(i)}} \vec{w_k} \tag{3.5}$$

Finally, the sense is chosen following Equation (3.6).

$$s = \arg\max_{s_i} cos(\vec{c^{(i)}}, \vec{s_i}) \tag{3.6}$$

Note that the context vector is changed according to individual senses when the similarity between the context and sense is measured.

Figure 3.4 shows an example to obtain a relevant word set $WR^{(i)}$. In this example, the target word is "argument" that have four senses, and CWS and T_r are set to 5 and 3, respectively. The bottom table shows the cosine similarity between the word vector of each word in a context and the sense vector of each sense. The values in bold indicate the three highest *RelevantScore* for each sense, and these words are chosen as the relevant word set, shown in the bottom of Figure 3.4.

Sentence:

While using those methods, values passed to those variables are called **arguments**.

Sense of *argument*:

s_1	a fact or assertion offered as evidence that something is true.
s_2	a reference or value that is passed to a function, procedure, subroutine, command, or program.
s_3	a summary of the subject or plot of a literary work or play or movie.
s_4	a contentious speech act; a dispute where there is strong disagreement.

Relevance score:

	method	value	pass	variable	call
s_1	**0.7**	0.1	**0.5**	**0.6**	0.3
s_2	0.5	**0.5**	**0.8**	**0.9**	0.4
s_3	0.3	**0.5**	0.2	**0.4**	**0.4**
s_4	**0.5**	**0.8**	**0.6**	0.3	0.4

$WR^{(1)} = \{$ method, pass, variable $\}$
$WR^{(2)} = \{$ value, pass, variable $\}$
$WR^{(3)} = \{$ value, variable, call $\}$
$WR^{(4)} = \{$ method, value, pass $\}$

Figure 3.4: Example of relevant word set

3.3 Collocation Based WSD

Unlike the HRWE method, this method determines the sense by only looking at a collocation, i.e. idiomatic phrase including a target word.

3.3.1 Collocation WSD Rule

Collocation WSD rule is defined in the following form:

$$\textit{collocation} \rightarrow \text{sense} = s_i \tag{3.7}$$

It means: when *collocation* appears in an input sentence, s_i is chosen as the sense of the target word.

Two types of the collocation are considered in this study. The first one is a word collocation that is a sequence of words including the target word.

$$
\begin{array}{rcl}
w_{i-2}\ w_{i-1}\ \mathbf{w} & \rightarrow & \text{sense}{=}s_i \\
w_{i-1}\ \mathbf{w} & \rightarrow & \text{sense}{=}s_i \\
w_{i-1}\ \mathbf{w}\ w_{i+1} & \rightarrow & \text{sense}{=}s_i \\
\mathbf{w}\ w_{i+1} & \rightarrow & \text{sense}{=}s_i \\
\mathbf{w}\ w_{i+1}\ w_{i+2} & \rightarrow & \text{sense}{=}s_i
\end{array}
$$

Figure 3.5: Template of word collocation WSD rule

$$
\begin{array}{rcl}
\mathbf{w}\text{ - }rel\text{ - }w_c & \rightarrow & \text{sense}{=}s_i \\
w_c\text{ - }rel\text{ - }\mathbf{w} & \rightarrow & \text{sense}{=}s_i
\end{array}
$$

Figure 3.6: Template of dependency collocation WSD rule

Five types of word collocation rule are defined as in Figure 3.5. **w** stands for the target word, while w_{i-2}, w_{i-1}, w_{i+1}, and w_{i+2} stand for words just before or after the target word. The suffix denotes a relative position of a word.

The second collocation is a dependency collocation, which is defined as a pair of words under a certain syntactic dependency. A syntactic dependency is a relation between two components in a sentence with one word being the governor and the other being the dependent of the relation[1]. Examples of the relation are "subject", "object", "modifier" and so on. Figure 3.6 shows the precise definition of the rule. **w** is the target word, while w_c is a word in a context that is under the dependency relation *rel* with **w**. It is well-known that a sense of a target word is strongly dependent to another word that has a dependency relation with the target word. That is the reason why the dependency collocation WSD rule is introduced in our system.

The use of the collocation WSD rule is quite simple. Figure 3.7 shows some usages of the collocation rules. In the example (a), the rule means that it chooses the sense s_1 when the target word "bank" appears in the collocation "bank robbery". When the context in (a) is given, the collocation in the rule is found or hit. Therefore, the rule determines the sense of the "bank" as s_1. Similarly, in the example (b), the sense of the target word "air" is chosen by simple matching of the collocation.

[1]This sentence is quoted from https://webanno.github.io/webanno/use-case-gallery/dependency-parsing/

(a)

collocation WSD rule:	bank robbery → sense=s_1(Financial institutions)
context:	during a **bank** robbery if robber has taken the bait money
chosen sense:	s_1

(b)

collocation WSD rule:	earth's atmosphere → sense=s_2(Air)
context:	the spacecraft disintegrated as it entered the Earth's **atmosphere**
chosen sense:	s_2

Figure 3.7: Example of usage of collocation WSD rule

3.3.2 Construction of Collocation WSD Rule

Collocation WSD rules are automatically acquired from a raw corpus. Figure 3.8 shows overall procedures.

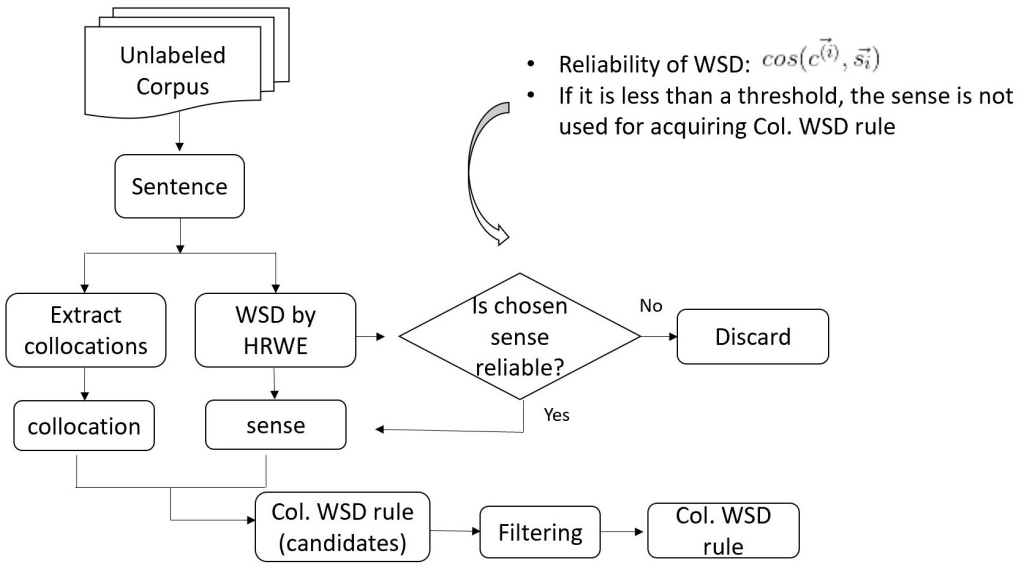


Figure 3.8: Flowchart of acquisition of collocation WSD rule.

First, for each sentence in an unlabeled corpus, the HRWE method determines a sense of a target word. If the chosen sense is reliable enough, the

sentence is used to obtain candidates of collocation WSD rules. The reliability of the disambiguated sense s_i is defined as the cosine similarity between the context vector and sense vector as shown in Equation (3.8).

$$reliability = cos(\vec{c^{(i)}}, \vec{s_i}) \tag{3.8}$$

If it is less than the threshold T_{wsd}, the sentence is just ignored.

Next, candidates of collocation WSD rules are generated by applying rule templates shown in Figure 3.5 and 3.6. For example, from the sentence "they were always getting into arguments about politics", where the HRWE determines the sense of "argument" as s_1, the candidates of the rules in Figure 3.9 are obtained. The first five rules are word collocation WSD rules, while the rest are dependency collocation WSD rules, which are derived from the dependency tree shown in Figure 3.10. Stanford Parser[2] is used to analyze dependency relations in this study.

[Sentence] they were always getting into arguments about politics

getting into **argument**	→	sense=s_1
into **argument**	→	sense=s_1
into **argument** about	→	sense=s_1
argument about	→	sense=s_1
argument about politics	→	sense=s_1
getting - *obj* - **arugment**	→	sense=s_1
argument - *case* - into	→	sense=s_1
argument - *nmod* - politics	→	sense=s_1

Figure 3.9: Obtained rules from example sentence

[2]https://nlp.stanford.edu/software/lex-parser.shtml

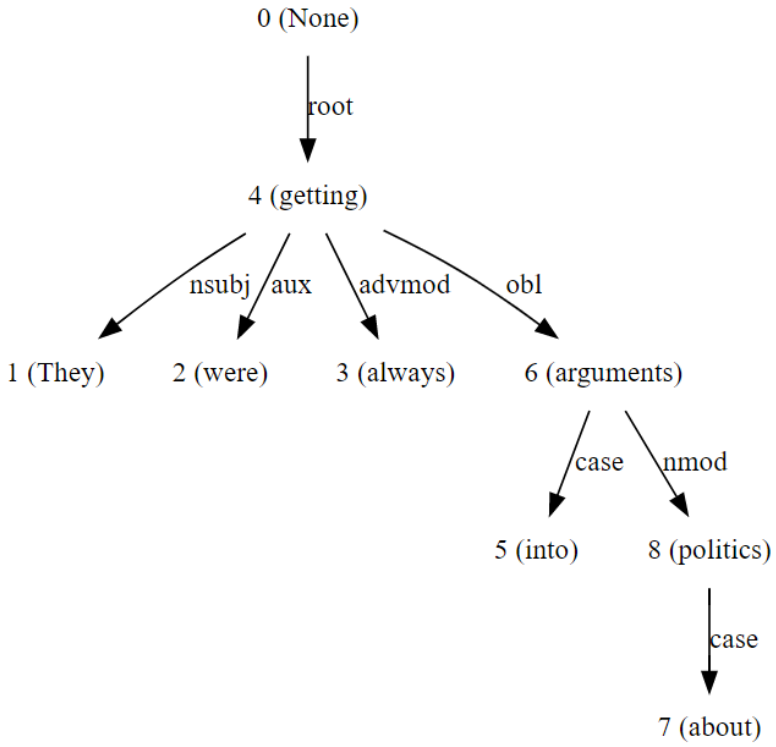


Figure 3.10: Dependency tree of example sentence

3.3.3 Filtering Collocation WSD Rule

After obtaining the candidates of the collocation WSD rules, inaccurate ones are filtered out. We apply the following three filtering procedures.

- **Stop word**
  The collocation consisting of only the target word and function words may not strongly associate with any senses. For example, the following rules determine a sense of the target word "play" or "argument" by looking up the collocation "play a" or "the argument". However, any senses of the target word can be appeared in such the collocation.

  $$\begin{array}{lcl} \textbf{play}\ \text{a} & \rightarrow & \text{sense}{=}s_2 \\ \text{the}\ \textbf{argument} & \rightarrow & \text{sense}{=}s_1 \end{array}$$

  Therefore, rules including collocations consisting of only function words (except for a target word) are discarded. We have prepared 28 function

word templates for this filtering. Figure 3.11 shows the list of them. In each template, **w** stands for a target word.

a **w**	**w** a	an **w**	any **w**	at **w**	**w** at
be **w**	**w** be	**w** but	for **w**	**w** for	**w** he
w his	**w** i	i **w**	in **w**	**w** in	it **w**
w it	or **w**	**w** or	that **w**	**w** that	the **w**
w the	the **w** the	this **w**	**w** this	**w** would	

Figure 3.11: Function word templates for filtering

- **Infrequent collocation**
  If the frequency of a collocation in a corpus is small, a rule might be unreliable. Therefore, rules are removed if the number of the collocation is less than the threshold T_{fre}.

- **Reliability**
  Obviously, not all rules are effective to choose a correct sense. Several rules are even inconsistent when the same collocation determines different senses such as "$col \rightarrow sense = s_1$" and "$col \rightarrow sense = s_2$". Therefore, the reliability score of the rule is defined as

$$score(col \rightarrow s_i) = \frac{f(col, s_i)}{\sum_i f(col, s_i)} \tag{3.9}$$

  , where $f(col, s_i)$ is the frequency of sentences including the collocation col and the sense s_i. Basically, this score means the precision of WSD when the sense is determined by the rule. If $score(col \rightarrow s_i)$ is less than T_{sco}, the rules are removed.

After applying these three filtering modules, the final set of collocation WSD rules is obtained.

The filtering process involves three parameters: T_{wsd} (the reliability of WSD), T_{fre} (the frequency of the collocation) and T_{sco} (the score of the rule). These parameters are empirically determined in our preliminary experiment. That is, the parameters are changed and the best ones are chosen by the evaluation of experimental results. T_{wsd} and T_{sco} are changed from 0.4 to 0.8 by a step of 0.05. T_{fre} is changed from 2 to 6 by a step of 1.

3.3.4 Summary

The characteristics of our proposed method can be summarized as follows.

- **Fully unsupervised WSD**
  Both the collocation WSD rule and the HRWE method are completely unsupervised methods. The implementation of this algorithm does not rely on any sense annotated corpora.

- **Use of both collocation feature and contextual feature**
  In the Basile's method, only contextual feature are used. On the other hand, in the method in this paper, both contextual feature and collocation feature are used. We believe that it is important to use different types of features for WSD to improve the performance.

- **Improvement of the context vector**
  The advanced approach to create the context vector is presented. It can filter out the information of noisy words from the context vector, causing the improvement of the quality of it.

Chapter 4

Evaluation

This chapter reports the experiments to evaluate our proposed method. Section 4.1 explains experimental setting, such as a dataset, evaluation criteria and so on. Section 4.2 reports preliminary experiments, that investigate the choice of the pre-training word embedding and the effectiveness of the gloss expansion. We present our experimental results in Section 4.3 to verify the effectiveness of our proposed method. Section 4.4 focuses on the experimental results of collocation WSD rules. Finally, Section 4.5 discusses influence of the context window size on WSD performance.

4.1 Experimental Setting

4.1.1 WordNet

As already explained, WordNet is used as a dictionary or a sense inventory in this study. WordNet is a commonly used lexical resource in NLP, which is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations[1]. Figure 4.1 shows a user interface of WordNet on Web[2]. Each sense entry includes a synset (set of synonyms), definition of a sense, and short example sentences. In this figure, we search for the word "drink" in the user interface of WordNet, and the system returned 5 senses of a noun and 5 senses of a verb. The text in brackets is the definition or gloss of a sense. Italics represent an example sentence of a corresponding sense. The underlined words represent words that belong to the same synset.

[1]Explanation of WordNet is quoted from https://wordnet.princeton.edu/
[2]http://wordnetweb.princeton.edu/perl/webwn

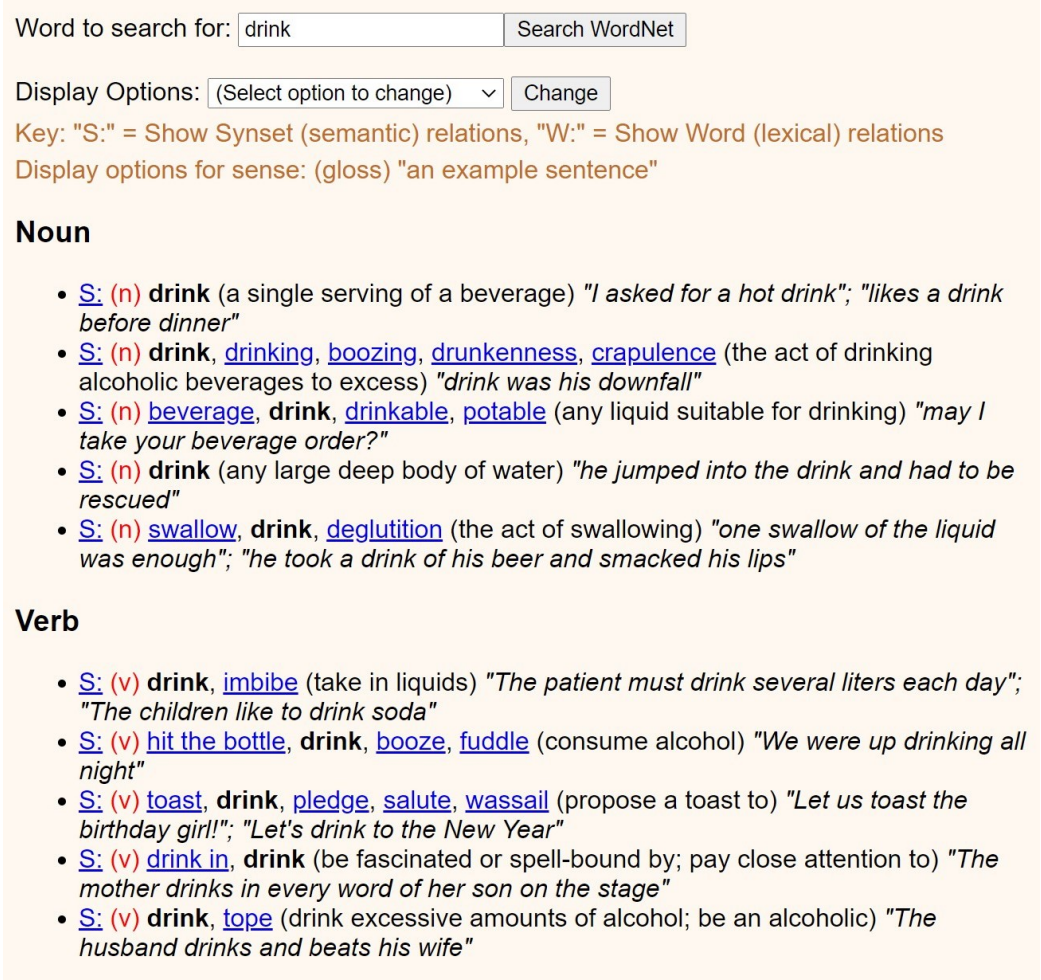


Figure 4.1: Example of Search on WordNet

4.1.2 Data

The dataset of Senseval-3 English lexical sample task is used to evaluate the performance of WSD of the proposed systems. It consists of instances (sentences or paragraphs including the target word) annotated with gold senses for several target verbs, nouns, and adjectives. The statistics of the dataset is shown in Table 4.1.

Table 4.2 shows all target words in the test data.

Table 4.1: Dataset of Senseval-3 English lexical sample task.

POS	# of words	ave.# of instances
Verb	27	53.1
Noun	17	78.5
Adjective	4	28.2
Total	48	59.8

Table 4.2: List of target words

Verb	activate, add, appear, ask, begin, climb, eat, encounter, hear, lose, mean, miss, play, produce, provide, receive, remain, rule, smell, suspend, talk, treat, use, wash, watch, win, write
Noun	argument, arm, bank, degree, difference, difficulty, disc, image, interest, judgment, paper, party, performance, plan, shelter, sort, source
Adjective	different, hot, important, solid

Figure 4.2 shows an example of a test instance. In this instance, the target word is "activating" (activate) in bold. "Context" means a paragraph including the target word. A system is required to determine a sense of the target word in this given context. "Sense tag" means the gold sense (correct sense) of the target word.

As preprocessing, lemmatization is performed for the sentences in the context. It converts words in a conjugated form to a base form. We use NLTK (Natural Language Tool Kit) [3] as a lemmatizer.

Context
...
and continue to have an important role in **activating** laity for what are judged to be religious goals both personally and socially.
...
Sense_tag
activate.v 38201

Figure 4.2: An example of test instance

In the Senseval-3 data, the senses are defined by WordNet [14]. As for the sense inventory, glosses in WordNet 1.7.1 are used for nouns and adjectives,

while definition sentences in Wordsmyth[3] are used for verbs. Since the senses in WordNet are fine-grained and differences of some senses are too subtle, we define a set of coarse-grained senses by manually merging similar senses. Table 4.3 shows examples of original senses and merged senses. Since the sense "bank%1:21:00" and "bank%1:21:01" are very similar, they are merged. A new ID "bank-c" is used to represent the unified sense. Similarly, "bank-a" and "bank-b" indicate the coarse-grained senses made by unifying several senses. The column "Unified sense ID" is left blank when an original sense is not unified with others such as "bank%1:04:00". The original and unified senses with their gloss sentences for all target words are reported in Appendix A.

Our WSD system is still developed to choose one of original (fine-grained) senses for a given target word, then the chosen sense is mapped to the corresponding unified (coarse-grained) sense. For example, when the system chooses either "bank%1:21:00" or "bank%1:21:01", it is converted to the new sense ID "bank-c". The performance of WSD is measured by comparing the predicted and gold coarse-grained senses.

Table 4.3: Example of sense unification

Target word	Original sense ID	Unified sense ID	Gloss sentence
bank	bank%1:04:00::		a flight maneuver; aircraft tips laterally about its longitudinal axis (especially in turning)
	bank%1:06:00::	bank-a	a building in which commercial banking is transacted
	bank%1:06:01::	bank-a	a container (usually with a slot in the top) for keeping money at home
	bank%1:14:00::	bank-a	a financial institution that accepts deposits and channels the money into lending activities
	bank%1:14:01::	bank-a	an arrangement of similar objects in a row or in tiers
	bank%1:17:00::	bank-b	a long ridge or pile
	bank%1:17:01::	bank-b	sloping land (especially the slope beside a body of water)
	bank%1:17:02::	bank-b	a slope in the turn of a road or track; the outside is higher than the inside in order to reduce the effects of centrifugal force
	bank%1:21:00::	bank-c	a supply or stock held in reserve for future use (especially in emergencies)
	bank%1:21:01::	bank-c	the funds held by a gambling house or the dealer in some gambling games

The average numbers of the senses per word in the original WordNet and

[3]http://www.wordsmyth.net/

our coarse sense set are shown in Table 4.4.

Table 4.4: Average number of senses

POS	WordNet	Our coarse sense
Verb	6.78	4.07
Noun	5.71	3.65
Adjective	11.5	3.00

A large unlabeled corpus is required to mine the collocation WSD rules. We suppose that a news corpus might be better than an online corpus, since news sentences are written in a firm style and grammatically correct. On the other hand, there are a lot of grammatically incorrect sentences on Web. In this experiment, 200,000 English news sentences from the Leipzig corpus[4] are used. Some preprocessing are performed on the corpus, such as lemmatization, and punctuation removal.

4.1.3 Evaluation criteria

Accuracy is a common evaluation criteria to evaluate WSD performance. It is defined a proportion of the correct judgement to the total number of a test data as shown in Equation (4.1). Precision is another criteria, which is defined as a proportion of the correct judgement to the cases where a WSD system can choose a (correct or incorrect) sense as in Equation (4.2).

$$\text{Accuracy} = \frac{\text{Number of data where a correct sense is chosen}}{\text{Number of all data}} \tag{4.1}$$

$$\text{Precision} = \frac{\text{Number of data that a system can determine a correct sense}}{\text{Number of data that a system can determine a sense}} \tag{4.2}$$

If a WSD system can always choose a senese for a given target instances, the accuracy and precision are the same. Our proposed method and Basile's method are such WSD systems. However, we also evaluate a WSD system using only collocation WSD rules. It fails to determine a sense when no rule is hit for a given test data. The precision is an appropriate evaluation criteria for such a system. As a result, we choose the precision as a major evaluation criteria in this experiment.

[4]https://wortschatz.uni-leipzig.de/en/download/english

4.2 Preliminary experiment

Two preliminary experiments are conducted. The first one is to compare word embedding models to be used in our proposed model. The results will be reported in Subsection 4.2.1. The second one is to verify the effectiveness of the gloss expansion, which will be discussed in Subsection 4.2.2.

4.2.1 Comparison of word embedding models

To construct the context and sense vectors, three pre-trained word embeddings are used: word embedding pre-trained by the Skip-gram model from Google News corpus[5], Glove[6][18], and BERT[7][4]. Since word embedding in BERT is dynamic, i.e. sentence-dependent, we expect that it is good to produce abstract vector representation of a context and sense. In our method, word embedding is used to construct the context vector and the gloss vector. They are formed by averaging word vectors in a context or a gloss sentence. That is, words are treated as a bag-of-words, where an order of words in a sentence is ignored. Therefore, the Skip-gram and Glove can be straightforwardly used in our method. When using BERT, the situation is a slightly different. BERT accepts a sentence, i.e. ordered word sequence, as an input. In this study, a textual fragment containing a target word is treated as a sentence in creation of the context vector, while a gloss sentence is given to BERT in creation of the gloss vector. The pre-trained BERT produces vectors for each word in a given sentence, and these vectors are used to construct the context and gloss vectors. In this experiment, we use BERT-Base from the google research repository[8].

Another issue on BERT is what layers should be used as word embedding. BERT are based on stacked transformers, multiple layers of transformers in other words. The BERT-Base model we use has 12 hidden layers. Each layer can produce word embedding. Therefore, there are several choices what layers to be used. Jawahar et al. discussed the difference between every layer for several NLP tasks [9]. Table 4.5 shows the performance of 10 NLP tasks when each of 12 BERT layer is used. Ten tasks are divided into three groups: surface task, syntactic task, and semantic task. Surface tasks are sentence length (SentLen) and the presence of words in the sentence (WC). Syntactic tasks are sensitivity to word order (BShift), the depth of the syntactic tree (TreeDepth) and the sequence of top-level constituents in the

[5]https://code.google.com/archive/p/word2vec/
[6]https://nlp.stanford.edu/projects/glove/
[7]https://github.com/google-research/bert
[8]https://github.com/google-research/bert

Table 4.6: Comparison of word embedding.

Type	Precision		
	Verb	Noun	Adjective
Skip-gram	**0.544**	**0.506**	**0.560**
Glove	0.529	0.484	0.468
BERT	0.424	0.495	0.504

syntax tree (TopConst). Semantic tasks are the tense (Tense), the subject and direct object number in the main clause (SubjNum and ObjNum), the sensitivity to random replacement of a noun/verb (SOMO) and the random swapping of coordinated clausal conjuncts (CoordInv). The value within the parentheses corresponds to the difference in performance of fine-tuned vs. not fine-tuned (only pre-trained) BERT. The result shows deep hidden layers perform better on semantic tasks, while shallow hidden layers learn more syntactic information. Since WSD is a semantic-related task, the output of last layer is used in our research.

Layer	SentLen (Surface)	WC (Surface)	TreeDepth (Syntactic)	TopConst (Syntactic)	BShift (Syntactic)	Tense (Semantic)	SubjNum (Semantic)	ObjNum (Semantic)	SOMO (Semantic)	CoordInv (Semantic)
1	93.9 (2.0)	24.9 (24.8)	35.9 (6.1)	63.6 (9.0)	50.3 (0.3)	82.2 (18.4)	77.6 (10.2)	76.7 (26.3)	49.9 (-0.1)	53.9 (3.9)
2	95.9 (3.4)	65.0 (64.8)	40.6 (11.3)	71.3 (16.1)	55.8 (5.8)	85.9 (23.5)	82.5 (15.3)	80.6 (17.1)	53.8 (4.4)	58.5 (8.5)
3	**96.2 (3.9)**	66.5 (66.0)	39.7 (10.4)	71.5 (18.5)	64.9 (14.9)	86.6 (23.8)	82.0 (14.6)	80.3 (16.6)	55.8 (5.9)	59.3 (9.3)
4	94.2 (2.3)	**69.8 (69.6)**	39.4 (10.8)	71.3 (18.3)	74.4 (24.5)	87.6 (25.2)	81.9 (15.0)	81.4 (19.1)	59.0 (8.5)	58.1 (8.1)
5	92.0 (0.5)	69.2 (69.0)	40.6 (11.8)	81.3 (30.8)	81.4 (31.4)	89.5 (26.7)	85.8 (19.4)	81.2 (18.6)	60.2 (10.3)	64.1 (14.1)
6	88.4 (-3.0)	63.5 (63.4)	**41.3 (13.0)**	83.3 (36.6)	82.9 (32.9)	89.8 (27.6)	**88.1 (21.9)**	82.0 (20.1)	60.7 (10.2)	71.1 (21.2)
7	83.7 (-7.7)	56.9 (56.7)	40.1 (12.0)	**84.1 (39.5)**	83.0 (32.9)	89.9 (27.5)	87.4 (22.2)	**82.2 (21.1)**	61.6 (11.7)	74.8 (24.9)
8	82.9 (-8.1)	51.1 (51.0)	39.2 (10.3)	84.0 (39.5)	83.9 (33.9)	89.9 (27.6)	87.5 (22.2)	81.2 (19.7)	62.1 (12.2)	76.4 (26.4)
9	80.1 (-11.1)	47.9 (47.8)	38.5 (10.8)	83.1 (39.8)	**87.0 (37.1)**	**90.0 (28.0)**	87.6 (22.9)	81.8 (20.5)	63.4 (13.4)	**78.7 (28.9)**
10	77.0 (-14.0)	43.4 (43.2)	38.1 (9.9)	81.7 (39.8)	86.7 (36.7)	89.7 (27.6)	87.1 (22.6)	80.5 (19.9)	63.3 (12.7)	78.4 (28.1)
11	73.9 (-17.0)	42.8 (42.7)	36.3 (7.9)	80.3 (39.1)	86.8 (36.8)	89.9 (27.8)	85.7 (21.9)	78.9 (18.6)	64.4 (14.5)	77.6 (27.9)
12	69.5 (-21.4)	49.1 (49.0)	34.7 (6.9)	76.5 (37.2)	86.4 (36.4)	89.5 (27.7)	84.0 (20.2)	78.7 (18.4)	**65.2 (15.3)**	74.9 (25.4)

Table 4.5: Probing task performance for each BERT layer [9]

Table 4.6 shows the average precision for disambiguation of the test data using the Basile's method with different word embedding models. Here the context window size *CWS* is set to 10. Although BERT can generate contextual embedding, it is found that the performance of the BERT is not the best. It indicates that pre-trained BERT model may not be appropriate for WSD. Since the result of this experiment indicates that the Skip-gram model is the best, only the Skip-gram model is used in our experiments.

4.2.2 Evaluation of gloss expansion

A preliminary experiment is carried out to confirm the effectiveness of the gloss expansion. As explained in Subsection 3.2.1, in Basile's method, not

only gloss sentences but also glosses of its related words (hypernym, hyponym, and synonym) are used to make a sense vector. Table 4.7 shows the precision of WSD for nouns to compare the original Basile's method with and without the gloss expansion. Although Basile et al. reported that gloss expansion was effective [1], it is not true in our experiment using Senseval-3 dataset. Therefore, the gloss expansion is not performed in the rest of experiments.

Table 4.7: Evaluation of gloss expansion on Senseval-3

Model	Precision		
	Noun	Verb	Adjective
w/o expansion	0.505	0.542	0.560
with expansion	0.457	0.517	0.447

4.3 Results

The several ideas to improve the WSD performance are newly introduced in our proposed method: to construct the context vector with only contextual words that are highly related to the sense (HRWE), to acquire and combine the collocation WSD rules, to use two types of collocation WSD rules (word collocation and dependency collocation), and so on. In this section, the contribution of each component in our method is discussed through the comparison of different WSD models.

First, the HRWE method is evaluated. Table 4.8 reports the results of two WSD methods, the baseline that is equivalent to [1], and the WSD system using our proposed HRWE method only. Note that the collocation WSD rules are not used. The table shows the average precision for verbs, nouns, and adjective as well as the average precision of all POSs in the "All" column. The HRWE outperforms the baseline for nouns and verbs, but not for adjectives. However, the precision is improved by 3.2 point for all POSs. It indicates that our idea to select contextual words strongly associated with senses for the context embedding is effective.

Table 4.8: Comparison of Baseline and HRWE method

Method	Precision			
	Verb	Noun	Adj	All
Baseline	0.542	0.506	**0.560**	0.525
HRWE only	**0.583**	**0.534**	0.511	**0.557**

Table 4.9 shows the precision and applicability of the system using the collocation WSD rules only. The applicability is a proportion of test instances that can be disambiguated by a WSD system (by the collocation WSD rules in this case). Equation (4.3) shows the definition of it.

$$\text{Applicability} = \frac{\text{Number of data that a system can determine a sense}}{\text{Number of all data}} \tag{4.3}$$

The applicability of the rules is low, i.e. senses in many sentences cannot be determined. However, the rules tend to achieve the higher precision than the previous two systems, the baseline and HRWE only in Table 4.8, especially for nouns and adjectives. It is confirmed that we can obtain the disambiguation rules whose recall is low but precision is high as we aimed. Note that the applicability of all other WSD systems is 1, that is, senses of all target instances are determined.

Table 4.9: Results of the system using the collocation WSD rule only

Index	Verb	Noun	Adj	All
Precision	0.573	0.631	0.625	0.591
Applicability	36.4%	17.8%	11.3%	26.8%

Table 4.10 shows the performance of the systems integrating the baseline or HRWE with the word or dependency collocation WSD rules. The use of two different WSD systems can increase the precision. Therefore, it is confirmed that both words in a context (considered in the baseline or HRWE) and collocations (considered in the rules) can contribute to choose the appropriate sense. Comparing 4-th and 5-th or 7-th and 8-th rows, the contribution of two types of collocation WSD rules (word vs. dependency) are almost equivalent.

Table 4.10: Comparison of systems with or without collocation WSD rule

Method	Precision			
	Verb	Noun	Adj	All
Baseline	0.542	0.506	**0.560**	0.525
Baseline + word collocation	0.553	0.516	0.553	0.536
Baseline + dep. collocation	0.547	0.510	0.546	0.530
HRWE only	0.583	0.534	0.511	0.557
HRWE + word collocation	0.588	**0.545**	0.511	**0.565**
HRWE + dep. collocation	**0.589**	0.540	0.525	0.564

Finally, Table 4.11 shows the precision of three WSD systems: Baseline, HRWE only, and the HRWE and both word and dependency collocation WSD rules. The last one is our final system. It achieves the best performance for nouns, verbs and all POSs as indicated in bold. Its precision is 0.572, which is 4.7 point better than the baseline.

It is found that our HRWE and collocation WSD rules poorly perform for the disambiguation of adjectives. However, the number of target adjectives in the test data is rather small, i.e. only 4. We will evaluate our proposed method for more adjectives and investigate how our system can improve sense disambiguation of adjectives in future.

Table 4.11: Comparison of WSD methods

Method	Precision			
	Verb	Noun	Adj	All
Baseline	0.542	0.506	**0.560**	0.525
HRWE only	0.583	0.534	0.511	0.557
HRWE + word & dep. collocation	**0.594**	**0.552**	0.525	**0.572**

4.4 Evaluation of collocation WSD rules

The details of the acquisition of the collocation WSD rules are reported in this subsection.

Recall that there are three thresholds for rule acquisition: T_{wsd} (the reliability of WSD), T_{fre} (the frequency of the collocation), and T_{sco} (the score of the rule). As described in Subsection 3.3.3, these parameters are empirically determined for individual POSs by trial and error in test data. Table 4.12 shows the chosen values of three parameters for individual POSs.

Table 4.12: Parameters for acquisition of collocation WSD rule

	T_{wsd}	T_{fre}	T_{sco}
Verb	0.75	4	0.7
Noun	0.7	5	0.7
Adjective	0.7	4	0.7

Table 4.13 shows the number of candidates of rules and rules after the filtering. Around five hundred word collocation WSD rules and nine hundred dependency collocation WSD rules are finally obtained. It is found that most of the candidates are inaccurate and discarded by our filtering methods.

Table 4.13: Number of rules mined from raw corpus

	candidates	after filtering
word collocation rule	132,300	528
dependency collocation rule	120,170	379

Many rules are intuitively right to choose the correct sense. Figure 4.3 shows the examples of acquired rules.

bank robber → sense=s_3(financial institute)
running **arguments** → sense=s_2(parameter)
talk - advmod - speechify → sense=s_1(speech)
refute - obj - **argument** → sense=s_1(assertion)

Figure 4.3: Example of acquired collocation WSD rule

For example, the first rule shows the meaning of "bank" in the collocation "bank robber" is "financial institue". The second rule shows the meaning of "arguments" in collocation "running arguments" is "parameter". The third and fourth are dependency collocations, which including a dependency relationship between words. The third rule indicates that when "speechify" modifies the verb "talk", the sense of "talk" is s_1 (speech). The fourth rule indicates that when "argument" is an object of the verb "refute", its meaning is s_1 (assertion).[9]

In addition to improvement of the precision, another minor advantage of incorporating collocation based WSD method is to shorten the computational time of disambiguation process. Since the collocation WSD rules can be retrieved and stored beforehand, the sense can be determined by simple pattern matching in a test phase. Although no experiment is conducted to compare the processing time of our method and the baseline method, we believe that our method is faster because a part of instances in the test data can be determined by the rules instead of heavy calculation of the context and sense vectors. Empirical comparison of computational costs remains to be carried out in future work.

4.5 Discussion about context window size

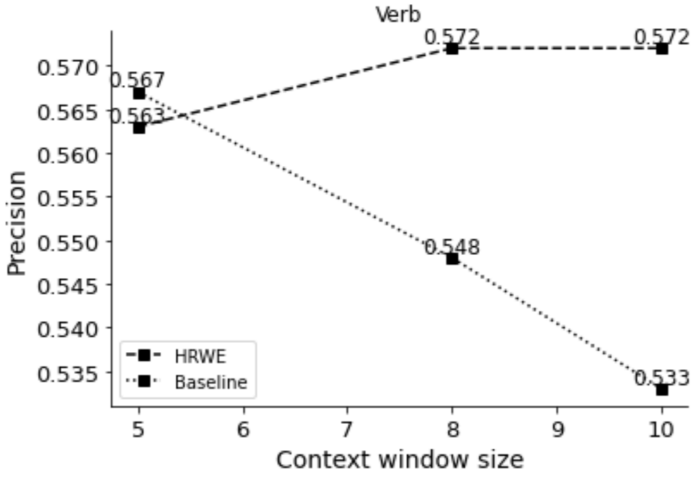
Next, influence of the context window size *CWS* on the WSD performance is investigated. *CWS* is changed to 5, 8, and 10 in the baseline and HRWE

[9]See also the sense definition in Figure 3.4.

method, then the WSD precision of these models are compared. Note that collocation WSD rules are not used in this experiment. Figure 4.4 (a) and (b) show the results for verb and noun, respectively.[10] The precision of our HRWE method is improved when *CWS* is increased, while that of the baseline is declined for both verbs and nouns. In the baseline method, when more context words are added to the context vector, words irrelevant to the correct sense are also added more. It results in spoiling the quality of the context vector. On the other hand, in the HRWE, not all but fixed number of highly related words are used to make the context vector. When the context window size is increased, words that are far from a target word but effective for WSD can be taken into account.

[10]A result of adjective is omitted since the number of target words in the test data is small.

(a) Verb



(b) Noun

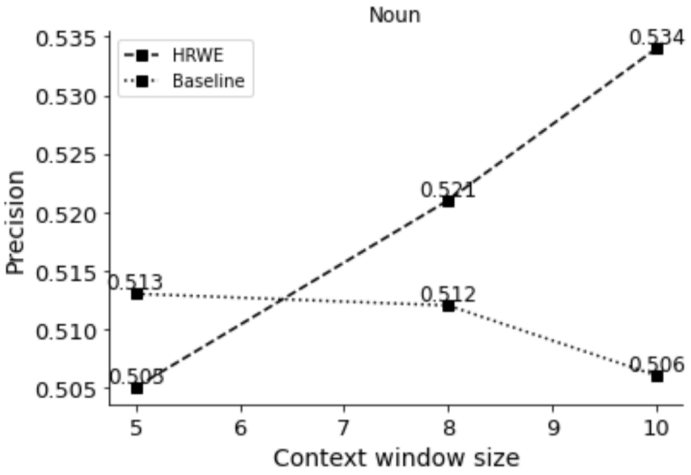


Figure 4.4: Precision of models with different context window sizes

Chapter 5

Conclusion

5.1 Concluding Remark

This paper proposed the novel unsupervised WSD system consisting of two methods. The first method was the method to determine the sense by looking up the collocation that strongly indicated the sense of the target word. Two types of the collocation WSD rules were acquired from a raw corpus, one is word collocation and the other is dependency collocation. The second method was the HRWE method that measured the similarity between the context and the gloss sentences, where noisy words were ignored in the construction of the context vector. The experimental results on Senseval-3 English lexical sample task dataset showed that our proposed method outperformed the previous work [1] by 4.7 point.

The contribution of the paper was summarized as follows. First, the collocation was newly integrated as another useful feature into the existing word embedding based method, which only considered words in the context. Ensemble of collocation based and word embedding based methods was effective to improve the precision of WSD. Another contribution was to refine how to make the context vector, where only highly related words were chosen to get better representation of the context.

5.2 Future Work

Due to the complexity of the task, WSD field has many directions worth exploring in the future. In our plan, more sophisticated methods to make the context and sense vectors will be explored. For example, it is worth investigating a method to use Sentence BERT [19] to obtain the vector representation of the sentences. We will explore the reason of the contextual

word embeddings generated by Deep Neural Networks (i.e. BERT) do not perform as well as traditional word embeddings like Skip-gram model on our method. Next, there is much room to improve the quality of collocation WSD rules. We need to explore more filtering methods to choose really good rules from candidates. In the experiment, we found that the collocation WSD rules of some target words are very effective but some are poor. It indicates that there exists two kinds of words: one is a word whose senses can be often determined by the collocation, the other is a word whose senses are independent to the collocation. If two types of the target words can be automatically distinguished and the WSD system based on the collocation WSD rules is used only for the former type, the overall performance of WSD will be improved. Another important line is to combine other unsupervised methods such as graph based ones with our HRWE method and collocation WSD rules. In addition, we need to assess why the gloss expansion was ineffective in our experiment.

Bibliography

[1] Pierpaolo Basile, Annalina Caputo, and Giovanni Semeraro. An enhanced Lesk word sense disambiguation algorithm through a distributional semantic model. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1591–1600, 2014.

[2] CE Bazell. Studies in linguistic analysis. special volume of the philological society, vii, 205 pp., 5 plates. oxford: Basil blackwell, 1957. 70s. *Bulletin of the School of Oriental and African Studies*, 22(1):182–184, 1959.

[3] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. "O'Reilly Media, Inc.", 2009.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[5] Miguel Ángel Ríos Gaona, Alexander Gelbukh, and Sivaji Bandyopadhyay. Web-based variant of the lesk approach to word sense disambiguation. In *2009 Eighth Mexican International Conference on Artificial Intelligence*, pages 103–107, 2009.

[6] Yuhang Guo, Wanxiang Che, Yuxuan Hu, Wei Zhang, and Ting Liu. Hit-ir-wsd: A wsd system for english lexical sample task. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 165–168, 2007.

[7] Zellig Harris. Mathematical structures of language. *Interscience tracts in pure and applied mathematics*, 1968.

[8] Luyao Huang, Chi Sun, Xipeng Qiu, and Xuanjing Huang. GlossBERT: BERT for word sense disambiguation with gloss knowledge. *arXiv preprint arXiv:1908.07245*, 2019.

[9] Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. What does bert learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, page 3651–3657, 2019.

[10] Cuong Anh Le and Akira Shimazu. High wsd accuracy using naive bayesian classifier with rich features. In *Proceedings of the 18th Pacific Asia Conference on Language, Information and Computation*, pages 105–114, 2004.

[11] Michael Lesk. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th annual international conference on Systems documentation*, pages 24–26, 1986.

[12] John C Mallery. Thinking about foreign policy: Finding an appropriate role for artificially intelligent computers. In *Master's thesis, MIT Political Science Department*, 1988.

[13] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26:3111–3119, 2013.

[14] George A Miller. WordNet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[15] George A Miller, Claudia Leacock, Randee Tengi, and Ross T Bunker. A semantic concordance. In *HUMAN LANGUAGE TECHNOLOGY: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993*, 1993.

[16] Roberto Navigli. Word sense disambiguation: A survey. *ACM computing surveys (CSUR)*, 41(2):1–69, 2009.

[17] Roberto Navigli and Mirella Lapata. An experimental study of graph connectivity for unsupervised word sense disambiguation. *IEEE transactions on pattern analysis and machine intelligence*, 32(4):678–692, 2009.

[18] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[19] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. *arXiv preprint arXiv:1908.10084*, 2019.

Appendix A

Sense Inventory

The following tables show the sense inventory of the target words used in the experiment. Each table shows the original senses and the unified senses (coarse-grained senses defined by us) in the same format of Table 4.3. Table A.1 - A.6, Table A.7 - A.11, and Table A.12 - A.14 show the senses of the verbs, nouns, and adjective, respectively. Note that the formats of the original sense ID are different for verbs (numerical ID such as "38201") and nouns/adjectives (ID such as "argument%1:09:00"), since the former is excerpted from Wordsmyth and the latter is excerpted from WordNet.

Table A.1: Original and unified senses of verb (1)

Target word	Original sense ID	Unified sense ID	Gloss sentence
activate	38201		to initiate action in; make active
	38202		in chemistry, to make more reactive, as by heating
	38203		to assign (a military unit) to active status
	38204		in physics, to cause radioactive properties in (a substance)
	38205		to cause decomposition in (sewage) by aerating
add	42601		to combine (something) with something else, often to increase the amount or number of the latter
	42602	add-a	to find the total of (often fol. by up)
	42605	add-a	to make the correct total or expected result (fol. by up)
	42603		to say or write beyond what has been said or written
	42604		to perform the mathematical operation of addition
	42606		to increase (fol. by to)

Table A.2: Original and unified senses of verb (2)

Target word	Original sense ID	Unified sense ID	Gloss sentence
appear	190901		to come into view; become visible
	190902		to seem
	190903		to come before the public, as a book or performer
ask	238101	ask-a	to put a question to
	238105	ask-a	to question; inquire
	238102	ask-b	to request of
	238104	ask-b	to invite
	238106	ask-b	to request or seek (usu. fol. by for)
	238103		to demand or expect
begin	369201		to perform the first step in a process; start
	369202		to come into being
	369203		to perform the first step of (something); start
	369204		to cause to come into being
climb	770001	climb-a	to move upward; go towards the top; ascend
	770002	climb-a	to slope upward
	770005	climb-a	to go up; ascend
	770003		to twist around and up a tall support
	770004		to strive to become more important, wealthier, or more successful, or to become so
eat	1297001	eat-a	to consume (food) through the mouth
	1297006	eat-a	to partake of food
	1297002	eat-b	to destroy through wearing away; corrode
	1297007	eat-b	to corrode
	1297003		to ravage or consume in the manner of eating
	1297004		to bother or disturb
	1297005		(informal) to bear the cost of
encounter	1353101	encounter-a	to meet or come upon, esp. suddenly or by chance
	1353103	encounter-a	to meet with, or come up against, esp. unexpectedly
	1353102		to meet or confront in battle or conflict
	1353104		to meet, esp. in conflict or unexpectedly
hear	1892101	hear-a	to perceive with the ears
	1892103	hear-a	to listen to carefully
	1892105	hear-a	to have the ability to perceive sound
	1892102	hear-b	to be informed about; learn
	1892104	hear-b	to give formal audience to, esp. in a court of law
	1892106	hear-b	to receive information or greetings from another
	1892107	hear-b	to listen with agreement or consent (usu. fol. by of)

Table A.3: Original and unified senses of verb (3)

Target word	Original sense ID	Unified sense ID	Gloss sentence
lose	2439901	lose-a	to no longer possess; be unable to find; misplace
	2439902	lose-a	to fail to keep possession of
	2439904	lose-a	to fail to maintain; be unable to keep
	2439905	lose-a	to suffer the loss of through death
	2439903	lose-b	to fail to win
	2439906	lose-b	to fail to use or take advantage of; waste
	2439908	lose-b	to experience defeat or loss
	2439907		to go astray from
	2439909		to diminish the effectiveness in a particular way
mean	2555501	mean-a	to have as a goal or purpose; intend
	2555502	mean-a	to intend to denote or express
	2555507	mean-a	to have intentions or be disposed
	2555503		of words, to signify
	2555504	mean-b	to intend for a particular purpose or end
	2555505	mean-b	to cause as a result
	2555506	mean-b	to have a specified degree of significance or importance
miss	2644301	miss-a	to fail to hit, catch, reach, cross, or in any way touch or contact (a particular object)
	2644302	miss-a	to fail to see, hear, understand, or otherwise acknowledge
	2644303	miss-a	to fail to perform, attend, or otherwise experience
	2644307	miss-a	to fail to hit, catch, or otherwise touch something such as a target, ball, or other object
	2644304		to fail to achieve or attain
	2644305		to avoid, escape, or evade
	2644306		to feel sad or lonely in the absence of
	2644308		to fail; not succeed
play	3165210		to act the part of in a drama
	3165211		to act (a role) in real life
	3165212		to perform in (a place or places)
	3165213	play-a	to take part in (a game or contest)
	3165217	play-a	to engage in recreation; have fun
	3165218	play-a	to engage in a sport or game
	3165214	play-b	to make music with (an instrument)
	3165220	play-b	to make music with an instrument
	3165215		to manipulate for one's advantage (usu. fol. by off)
	3165216		to control (a hooked fish)
	3165219		to behave in a specified way
	3165221		to make a toy of another; use another without due regard for his or her feelings

Table A.4: Original and unified senses of verb (4)

Target word	Original sense ID	Unified sense ID	Gloss sentence
produce	3288301	produce-a	to bring into being; yield
	3288306	produce-a	to cause, create, or yield results, esp. the usual or expected results
	3288302	produce-b	to manufacture
	3288305	produce-b	to organize and present (a film, play, concert, or the like) for public entertainment
	3288303		to give birth to
	3288304		to bring forward into view or notice; present
provide	3313901	provide-a	to supply; furnish
	3313906	provide-a	to make an arrangement, agreement, or condition
	3313902	provide-b	to make available for use; afford
	3313905	provide-b	to supply necessities such as money (often fol. by for)
	3313903		to arrange or specify beforehand
	3313904		to take precautionary action (usu. fol. by for or against)
receive	3434801	receive-a	to get or take (something) that has been sent or offered
	3434806	receive-a	to accept or get something
	3434802		to accept (something) that has been bestowed
	3434803	receive-b	to welcome
	3434807	receive-b	to extend hospitality to guests
	3434804		to undergo; experience
	3434805		to find out about
	3434808		to pick up signals, as on a radio or television
	3434809		in football, to play in the position of one designated to catch a forward pass
remain	3477801	remain-a	to continue without a change in quality or state
	3477803	remain-a	to be left, as still to be done
	3477802		to stay or be left in the same place after others have gone
rule	3597906	rule-a	to exert authority over; govern
	3597907	rule-a	to have superiority over, within a particular field or area
	3597911	rule-a	to be pervasive or dominant
	3597908		to make evenly spaced parallel lines on (a piece of paper or other surface)
	3597910		to make a specific decision or ruling, as in a court of law

Table A.5: Original and unified senses of verb (5)

Target word	Original sense ID	Unified sense ID	Gloss sentence
smell	3893501	smell-a	to perceive the odor of by means of the nose
	3893505	smell-a	to have or give off an odor or fragrance
	3893507	smell-a	to have or give off an unpleasant odor; stink
	3893508	smell-a	to have a lingering trace (usu. fol. by of)
	3893502	smell-b	to examine by using the sense of smell
	3893503	smell-b	to detect; discern
	3893509	smell-b	to investigate (usu. fol. by about or around)
suspend	4155301		to hang (something) from a higher position
	4155302	suspend-a	to cause to stop for a period of time
	4155303	suspend-a	to put off till later; defer
	4155304	suspend-a	to cause to be temporarily ineffective
	4155307	suspend-a	to cease activity for a period of time
	4155305		to exclude for disciplinary reasons
	4155306		to cause to remain motionless, undissolved, or unattached in a fluid medium such as air or water
talk	4198501	talk-a	to communicate through spoken words; discuss
	4198502	talk-a	to gossip
	4198503	talk-a	to chatter idly or incessantly
	4198506	talk-a	to articulate in words
	4198507	talk-a	to speak (a particular language or dialect)
	4198504		to give a speech; lecture
	4198505		to disclose confidential or secret information
	4198508		to discuss
	4198509		to influence; convince
treat	4380101	treat-a	to behave toward (someone) in a particular way
	4380102	treat-a	to deal with or represent in a particular way
	4380103	treat-a	to discuss in speech or writing
	4380104	treat-b	to relieve or cure (a disease or illness)
	4380105	treat-b	to give medical attention to
	4380106	treat-c	to offer (food, drink, or entertainment) to at one's own expense
	4380109	treat-c	to take responsibility for the cost of providing food, drink, or entertainment to another
	4380107		to act upon in order to achieve a desired result
	4380108		to deal with a subject, topic, or theme in speech or writing (often fol. by of)
use	4530701		to bring into service; employ, esp. habitually
	4530702	use-a	to expend; consume
	4530704	use-a	to partake of (drugs)
	4530703		to employ for selfish motives; exploit
	4530705		used in the past tense in order to show a former habitual practice or state (fol. by to)

Table A.6: Original and unified senses of verb (6)

Target word	Original sense ID	Unified sense ID	Gloss sentence
wash	4636101	wash-a	to make clean by immersing in or applying water or other liquid, esp. if soap is also used
	4636102	wash-a	to remove (dirt or other matter) by immersing in water or other liquid, esp. if soap is also used
	4636107	wash-a	to clean or bathe oneself
	4636108	wash-a	to clean something in or with water or other liquid
	4636103	wash-b	to transport by means of a moving liquid, esp. water
	4636104	wash-b	to erode or destroy by the action of moving water
	4636110	wash-b	to be carried by the action of moving water
	4636111	wash-b	to be removed or worn down by the action of moving water (often fol. by away)
	4636112	wash-b	to flow over; rush against
	4636105		to make wet; moisten; drench
	4636106		to rid of guilt or impurity
	4636109		to be capable of being cleaned in or with water without shrinking or fading
watch	4640501	watch-a	to look closely or with uninterrupted attention
	4640502	watch-a	to look or wait in alert expectation (usu. fol. by for)
	4640507	watch-a	to look at closely or with uninterrupted attention
	4640503	watch-b	to keep a vigil, esp. through the night
	4640508	watch-b	to guard or tend attentively
	4640509	watch-b	to stay informed about or aware of
	4640504		to be careful or alert
win	4711401	win-a	to be victorious in a competition
	4711403	win-a	to gain victory in
	4711405	win-a	to capture in battle
	4711402		to gain success through effort or struggle
	4711404	win-b	to obtain through effort
	4711406	win-b	to gain (loyalty, sympathy, affection, or the like)
	4711407		to succeed in obtaining the support of
write	4753401	write-a	to form (letters, words, symbols, or characters) on a surface with a pen, pencil, typewriter, or other instrument
	4753404	write-a	to fill in the spaces of or cover with writing
	4753406	write-a	to form letters, words, symbols, or characters on a surface with a pen, pencil, typewriter, or other instrument
	4753402		to express or record by writing
	4753403	write-b	to author or compose
	4753407	write-b	to create written material as a job or profession
	4753405	write-c	to leave the evidence or signs of
	4753408	write-c	to communicate by sending letters

Table A.7: Original and unified senses of noun (1)

Target word	Original sense ID	Unified sense ID	Gloss sentence
argument	argument%1:09:00::		a variable in a logical or mathematical expression whose value determines the dependent variable; if f(x)=y, x is the independent variable
	argument%1:10:00::		a discussion in which reasons are advanced for and against some proposition or proposal
	argument%1:10:01::		a summary of the subject or plot of a literary work or play or movie
	argument%1:10:02::		a fact or assertion offered as evidence that something is true
	argument%1:10:03::		a dispute where there is strong disagreement
arm	arm%1:06:00::		the part of a garment that is attached at armhole and provides a cloth covering for the arm
	arm%1:06:01::		instrument used in fighting or hunting
	arm%1:06:02::		the part of an armchair or sofa that supports the elbow and forearm of a seated person
	arm%1:06:03::		any projection that is thought to resemble an arm
	arm%1:08:00::		a human limb; technically the part of the superior limb between the shoulder and the elbow but commonly used to refer to the whole superior limb
	arm%1:14:00::		an administrative division of some larger or more complex organization
bank	bank%1:04:00::		a flight maneuver; aircraft tips laterally about its longitudinal axis (especially in turning)
	bank%1:06:00::	bank-a	a building in which commercial banking is transacted
	bank%1:06:01::	bank-a	a container (usually with a slot in the top) for keeping money at home
	bank%1:14:00::	bank-a	a financial institution that accepts deposits and channels the money into lending activities
	bank%1:14:01::	bank-a	an arrangement of similar objects in a row or in tiers
	bank%1:17:00::	bank-b	a long ridge or pile
	bank%1:17:01::	bank-b	sloping land (especially the slope beside a body of water)
	bank%1:17:02::	bank-b	a slope in the turn of a road or track; the outside is higher than the inside in order to reduce the effects of centrifugal force
	bank%1:21:00::	bank-c	a supply or stock held in reserve for future use (especially in emergencies)
	bank%1:21:01::	bank-c	the funds held by a gambling house or the dealer in some gambling games

Table A.8: Original and unified senses of noun (2)

Target word	Original sense ID	Unified sense ID	Gloss sentence
degree	degree%1:07:00::	degree-a	a position on a scale of intensity or amount or quality
	degree%1:07:01::	degree-a	the seriousness of something (e.g., a burn or crime)
	degree%1:26:01::	degree-a	a specific identifiable position in a continuum or series or especially in a process
	degree%1:09:00::		the highest power of a term or variable
	degree%1:10:00::		an award conferred by a college or university signifying that the recipient has satisfactorily completed a course of study
	degree%1:23:00::		a measure for arcs and angles
	degree%1:23:03::		a unit of temperature on a specified scale
difference	difference%1:07:00::		the quality of being unlike or dissimilar
	difference%1:10:00::		a disagreement or argument about something important
	difference%1:11:00::		a variation that deviates from the standard or norm
	difference%1:23:00::		the number that remains after subtraction; the number that when added to the subtrahend gives the minuend
	difference%1:24:00::		a significant change
difficulty	difficulty%1:04:00::		an effort that is inconvenient
	difficulty%1:07:00::		the quality of being difficult
	difficulty%1:09:02::		a factor causing trouble in achieving a positive result or tending to produce a negative result
	difficulty%1:26:00::		a situation or condition almost beyond one's ability to deal with and requiring great effort to bear or overcome
disc	disc%1:06:00::		a thin flat circular plate
	disc%1:06:01::		sound recording consisting of a disc with continuous grooves; formerly used to reproduce music by rotating while a phonograph needle tracked in the grooves
	disc%1:06:03::		(computer science) a memory device consisting of a flat disk covered with a magnetic coating on which information is stored
	disc%1:25:00::		something with a round shape like a flat circular plate

Table A.9: Original and unified senses of noun (3)

Target word	Original sense ID	Unified sense ID	Gloss sentence
image	image%1:06:00::	image-a	a visual representation of an object or scene or person produced on a surface
	image%1:06:01::	image-a	a representation of a person (especially in the form of sculpture)
	image%1:07:00::	image-b	(Jungian psychology) a personal facade one presents to the world
	image%1:09:00::	image-b	an iconic mental representation
	image%1:09:02::	image-b	a standard or typical example
	image%1:10:00::		language used in a figurative or nonliteral sense
	image%1:18:00::		someone who closely resembles a famous person (especially an actor)
interest	interest%1:04:01::	interest-a	a subject or pursuit that occupies one's time and thoughts (usually pleasantly)
	interest%1:07:01::	interest-a	a reason for wanting something done
	interest%1:07:02::	interest-a	the power of attracting or holding one's interest (because it is unusual or exciting etc.)
	interest%1:09:00::	interest-a	a sense of concern with and curiosity about someone or something
	interest%1:14:00::	interest-a	(usually plural) a social group whose members control some field of activity and who have common aims
	interest%1:21:00::	interest-b	a fixed charge for borrowing money; usually a percentage of the amount borrowed
	interest%1:21:03::	interest-b	a right or legal share of something; a financial involvement with something
judgment	judgment%1:04:00::	judgment-a	(law) the determination by a court of competent jurisdiction on matters submitted to it
	judgment%1:04:02::	judgment-a	the act of judging or assessing a person or situation or event
	judgment%1:10:00::	judgment-a	the legal document stating the reasons for a judicial decision
	judgment%1:07:00::	judgment-b	the capacity to assess situations or circumstances shrewdly and to draw sound conclusions
	judgment%1:09:00::	judgment-b	the cognitive process of reaching a decision or drawing conclusions
	judgment%1:09:01::	judgment-b	ability to make good judgments
	judgment%1:09:04::	judgment-b	an opinion formed by judging something

Table A.10: Original and unified senses of noun (4)

Target word	Original sense ID	Unified sense ID	Gloss sentence
paper	paper%1:06:00::	paper-a	a newspaper as a physical object
	paper%1:10:03::	paper-a	a daily or weekly publication on folded sheets; contains news and articles and advertisements
	paper%1:14:00::	paper-a	a business firm that publishes newspapers
	paper%1:10:00::	paper-b	medium for written communication
	paper%1:27:00::	paper-b	a material made of cellulose pulp derived mainly from wood or rags or certain grasses
	paper%1:10:01::	paper-c	an essay (especially one written as an assignment)
	paper%1:10:02::	paper-c	a scholarly article describing the results of observations or stating hypotheses
party	party%1:11:00::	party-a	an occasion on which people can assemble for social interaction and entertainment
	party%1:14:02::	party-a	a band of people associated temporarily in some activity
	party%1:18:00::	party-a	a person involved in legal proceedings
	party%1:14:00::		a group of people gathered together for pleasure
	party%1:14:01::		an organization to gain political power
performance	performance%1:04:00::	performance-a	the act of performing; of doing something successfully; using knowledge as distinguished from merely possessing it
	performance%1:04:01::	performance-a	the act of presenting a play or a piece of music or other entertainment
	performance%1:10:00::	performance-a	a dramatic or musical entertainment
	performance%1:04:03::	performance-b	any recognized accomplishment
	performance%1:22:00::	performance-b	process or manner of functioning or operating
plan	plan%1:06:00::		scale drawing of a structure
	plan%1:09:00::	plan-a	a series of steps to be carried out or goals to be accomplished
	plan%1:09:01::	plan-a	an arrangement scheme

Table A.11: Original and unified senses of noun (5)

Target word	Original sense ID	Unified sense ID	Gloss sentence
shelter	shelter%1:06:00::	shelter-a	a structure that provides privacy and protection from danger
	shelter%1:06:01::	shelter-a	protective covering that provides protection from the weather
	shelter%1:21:00::	shelter-b	a way of organizing business to reduce the taxes it must pay
	shelter%1:26:00::	shelter-b	the condition of being protected
sort	sort%1:07:00::		an approximate definition or example
	sort%1:09:00::		a category of things distinguished by some common characteristic or quality
	sort%1:18:00::		a person of a particular character or nature
	sort%1:22:00::		an operation that segregates items into groups according to a specified criterion
source	source%1:06:00::		a facility where something is available
	source%1:09:00::		anything that provides inspiration for later work
	source%1:10:00::	source-a	a document (or organization) from which information is obtained
	source%1:10:01::	source-a	a publication (or a passage from a publication) that is referred to
	source%1:18:00::	source-a	someone who originates or causes or initiates something
	source%1:15:00::		the place where something begins, where it springs into being
	source%1:18:01::		a person who supplies information

Table A.12: Original and unified senses of adjective (1)

Target word	Original sense ID	Unified sense ID	Gloss sentence
different	different%3:00:00::		unlike in nature or quality or form or degree
	different%3:00:02::		not like; marked by dissimilarity
	different%5:00:00:other:00		distinctly separate from the first
	different%5:00:00:unusual:00		differing from all others; not ordinary
	different%5:00:01:other:00		distinct or separate

Table A.13: Original and unified senses of adjective (2)

Target word	Original sense ID	Unified sense ID	Gloss sentence
hot	hot%3:00:01::	hot-a	used of physical heat; having a high or higher than desirable temperature or giving off heat or feeling or causing a sensation of heat or burning
	hot%3:00:02::	hot-a	extended meanings; especially of psychological heat; marked by intensity or vehemence especially of passion or enthusiasm
	hot%5:00:00:active:01	hot-b	(informal) marked by excited activity
	hot%5:00:00:charged:00	hot-b	(electricity) charged or energized with electricity
	hot%5:00:00:fast:01	hot-b	very fast
	hot%5:00:00:fresh:01	hot-b	newly made
	hot%5:00:00:good:01	hot-b	very good; often used in the negative
	hot%5:00:00:illegal:00	hot-b	(informal) recently stolen or smuggled
	hot%5:00:00:lucky:00	hot-b	having or bringing unusually good luck
	hot%5:00:00:near:00	hot-b	of a seeker; near to the object sought
	hot%5:00:00:new:00	hot-b	newest or most recent
	hot%5:00:00:popular:00	hot-b	(informal) very popular or successful
	hot%5:00:00:pungent:00	hot-b	having a piquant burning taste of spices or peppers
	hot%5:00:00:radioactive:00	hot-b	having or dealing with dangerously high levels of radioactivity
	hot%5:00:00:sexy:00	hot-b	sexually excited or exciting
	hot%5:00:00:skilled:00	hot-b	(informal) performed or performing with unusually great skill and daring and energy
	hot%5:00:00:unpleasant:00	hot-b	very unpleasant or even dangerous
	hot%5:00:00:violent:00	hot-b	characterized by violent and forceful activity or movement; very intense
	hot%5:00:00:wanted:00	hot-b	wanted by the police
	hot%5:00:00:warm:03	hot-b	(color) bold and intense
	hot%5:00:02:fast:01	hot-b	capable of quick response and great speed
	hot%5:00:00:eager:00		having or showing great eagerness or enthusiasm

Table A.14: Original and unified senses of adjective (3)

Target word	Original sense ID	Unified sense ID	Gloss sentence
important	important%3:00:00::	important-a	of great significance or value
	important%3:00:02::	important-a	of extreme importance; vital to the resolution of a crisis
	important%3:00:04::	important-a	important in effect or meaning
	important%5:00:00:immodest:02	important-b	having or suggesting a consciousness of high position
	important%5:00:00:influential:00	important-b	having authority or ascendancy or influence
solid	solid%3:00:01::	solid-a	of definite shape and volume; firm; neither liquid nor gaseous
	solid%3:00:02::	solid-a	entirely of one substance with no holes inside
	solid%5:00:00:cubic:00	solid-b	having three dimensions
	solid%5:00:00:frozen:00	solid-b	turned into or covered with thick ice
	solid%5:00:00:good:01	solid-b	of good substantial quality
	solid%5:00:00:hard:01	solid-b	not soft or yielding to pressure
	solid%5:00:00:homogeneous:00	solid-b	of one substance or character throughout
	solid%5:00:00:honorable:00	solid-b	having high moral qualities
	solid%5:00:00:opaque:00	solid-b	incapable of being seen through
	solid%5:00:00:plain:02	solid-b	entirely of a single color throughout
	solid%5:00:00:sound:01	solid-b	of good quality and condition; solidly built
	solid%5:00:00:unbroken:02	solid-b	uninterrupted in space; having no gaps or breaks
	solid%5:00:00:undiversified:00	solid-b	acting together as a single undiversified whole
	solid%5:00:00:wholesome:00	solid-b	providing abundant nourishment