Exploiting WordNet for Wikipedia-Based Named Entity Disambiguation

Thuan K. Nguyen†* ‡ ‡ ‡
Tru H. Cao†* ‡ ‡ ‡

† University of Science – VNUHCM
‡ ‡ University of Technology – VNUHCM
* John von Neumann Institute – VNUHCM
thuan.nguyen@jvn.edu.vn, tru@cse.hcmut.edu.vn

Abstract

Entity disambiguation is an important problem in semantic analysis and natural language processing. In this paper, we propose an approach to employ features of the WordNet ontology in the task of disambiguating named entities to Wikipedia. Methods of enriching text with synonymous relations of words are explored. An analysis of the results from our experiments shows that the accuracy of the disambiguation process is improved on a dataset of named entities.

Keywords: Entity Linking, Word Sense Disambiguation, Synset, Semantic Annotation

1 Introduction

The rapid growth of World Wide Web has been exposing an abundant and constantly updated data source to users around the globe. However, due to the unstructured nature of this large data collection; a crucial task of Natural Language Processing (NLP) is to better understand the content of a document through only some important keywords or phrases in the document. Such keywords or phrases are called entities and can be divided into two types: named entities (such as people, organizations and locations) and abstract entities expressing general concepts. Named entities carry essential information to convey the content of a document and thus are the central subject in many recent studies.

A named entity can be mentioned, or referred to, in a text by multiple possible surface forms (case 1) and conversely, a surface form can refer to multiple possible candidate entities (case 2). The latter case is where we need to perform disambiguation in order to tell which entity is actually referred to. The ability to identify named entities and properly link them to a knowledge base (KB) has been established as an important task in several areas, including topic detection and tracking, machine translation, and information retrieval [1, 4, 5, 10, 11]. This process is well-known as Entity Linking [9] or Entity Disambiguation [13]. Its main goal is to map mentions of entities into their correct referent entities in a provided KB used as a candidate inventory.

An example of case 1 described above is when “the United States of America” is referred to as “the USA”, “the U.S.”, or “the States”. For case 2, an example is when we need to perform disambiguation in order to determine the mention “Texas” in a document refers to Texas as a state of the USA, or Texas as a Charlie Daniels single from the album Nightrider, etc… The output of the disambiguation process is the identifier (ID) of the KB entry to which the mention refers; or NIL if there is no such KB entry [9].

Thanks to its broad coverage of subjects as well as its well organized structure, Wikipedia has been widely used in NLP researches and applications. It can serve as a reliable ontology for the entity disambiguation problem [1] and hence is chosen as the candidate inventory in our approach. The disambiguation result is a corresponding Wikipedia article for each given mention in the text document, or NIL if there is no such article.

Previous works on named entity disambiguation have employed different approaches, including machine learning [4, 5, 11], vector space model and rule-based [1, 2, 3, 8] or graph-based [15]. Even though these approaches exploited diverse features and employed different models in named entity disambiguation, the lexical aspect of words in the text has not been explored. Based on the observation that a word sense might be represented by multiple surface forms in a
given context, we present a disambiguation approach that extends the one proposed in [2] by integrating WordNet features into the entity linking process. This idea, however, is not limited to our presented model and can be applied to other approaches.

In particular, by taking into account the synonymous relations of words, we expand the bag of words for a mention in the text with synsets of the words contained inside. The synset expansion is applied either to all words after a word sense disambiguation process, or to words that have only a single sense.

The rest of the paper is organized as follows. Section 2 presents a brief overview of related works on named entity disambiguation. Details of the WIN algorithm upon which our system is built as well as relevant concepts in WordNet are explained in Section 3.1 and Section 3.2, respectively, followed by our proposed approach in Section 3.3. Next, in Section 4, experiment results and analyses are presented. In section 5, we draw conclusions on the explored methods and discuss possible future works.

2 Related Works

A typical approach to named entity disambiguation involves the measuring of context compatibility between a mention and its candidate entities. The Wikify! system presented in [5] trains a Bayes classifier for each ambiguous mention using three words to the left and the right of outgoing links in Wikipedia articles, with their parts-of-speech, as contextual features. However, this method requires a large amount of processing effort because of the parsing of entire Wikipedia.

Bunescu and Pașca [8] propose a method that uses an SVM kernel to compare the lexical context around the ambiguous mention to that of its candidate entities, in combination with estimating correlation of the contextual words with the categories of the candidate entities.

Cucerzan [1] was the first to model interdependence among disambiguation decisions. In [1], the authors exploit semantic relatedness which is based on overlap in categories of entities referred to in the text.

Han and Sun [15] propose a method based on a referent graph that performs collective inference of KB entities referred to in a text. A referent graph is a weighted and undirected graph \( G = (E, V) \) where \( V \) contains all mentions in the text and all possible candidates of these mentions. A mention-entity edge is established between a mention and an entity, and weighted based on context similarity, or a combination of popularity and context similarity. An entity-entity edge is established between two entities and weighted using semantic relatedness between them. The local context similarity is calculated based on a bag of words model and the semantic relatedness is computed using the formula presented in [4].

In [2], the authors present an incremental approach where already disambiguated entities help in the process of disambiguating other remaining entities. The incremental mechanism of this approach is similar to human processing of mentions based on previously known ones. This approach is also a novel hybrid approach of combining heuristic rules and statistics compared to other methods discussed earlier.

Although these existing works have used different strategies for entity disambiguation, only the exact forms of words are considered. The measuring of the context compatibility between a mention and its candidate entities thus has yet considered the case of multiple surface forms for a word. Consequently, our approach aims to explore the effect of using additional surface forms for a word in measuring the context compatibility.

3 Proposed Approach

As previously mentioned, our proposed approach extends the original work in [2]. In the following self-contained Section 3.1, we provide an overview of the steps involved in this disambiguation approach. In Section 3.2, we present relevant WordNet concepts that are used in our work. Finally in Section 3.3 we present our proposed approach of exploiting WordNet synsets in the disambiguation process.

3.1 WIN Algorithm

In [2], the authors propose an incremental process to disambiguate named entities. The main idea of this approach is to use the identifiers of already disambiguated entities to help in the disambiguating of later entities.

An input text document is first passed through a preprocessing phase for entity recog-
inition and entity coreference resolution (identifying each group of mentions that actually refer to the same entity). The second phase is disambiguation where the system is required to link each mention recognized in the text to its corresponding entity in Wikipedia. The disambiguation phase consists of heuristics and statistics stages; we focus on the stage of statistics where the ranking of possible Wikipedia candidate articles is performed.

In the statistics stage, a bag-of-words representation of text is employed to compute cosine similarity between two types of vector: vector representing a mention in the text and vector representing a candidate Wikipedia article. The higher the cosine similarity is, the higher the rank that candidate receives. The candidate with highest rank is chosen as the “correct” mapping for this mention into Wikipedia.

Each mention in the text can appear in a certain context (in other words, a specific position in the text). The bag of words representing a mention exploits the following features extracted from the text:

1. Local words surrounding the mention: all tokens in a window of 55 tokens surrounding the examined mention.
2. Local words surrounding coreference mentions.
3. All mentions in the document.
4. Identifiers of already disambiguated mentions: in the case of Wikipedia KB, such an identifier is the title of an article page.

The bag of words representing a Wikipedia article of an entity includes the following features:

1. Title of the entity page.
2. Title of redirect pages.
3. Category labels: a list of categories that this entity belongs to.
4. Outgoing links: titles of the entity pages that these links point to are used.

Extracted features for a mention in a text or an article in Wikipedia are then put into a bag-of-words after a normalizing process.

3.2 WordNet lexical database

WordNet [6, 7] is a large lexical database for English organized in synonym sets (synsets). As of version 2.1, WordNet contains about 150,000 words placed in over 115,000 synsets. Nouns, verbs, adjectives and adverbs are grouped into sets of synonyms — words that share the same concept and can be used interchangeably in a certain context.

Each word in a text may be mapped to a word form in WordNet if existing in the ontology. All word forms or synonyms in the same synset share the same concept and can be represented by a WordNet sense. For instance, in the 4th sense of the noun “good”, the two word forms “good” and “commodity” express the identical meaning of articles of commerce. Thus these words constitute part of a synset with its corresponding sense identifier (sense ID).

Figure 1 shows a comparison of how Wikipedia (on the left side) and WordNet (on the right side) visually differentiate among possible entities or word senses for the mention “plant”. In the Wikipedia’s disambiguation page there is a list of links that point to articles representing different types of plant, such as “plant” as a living organism (a tree), or “plant” as a chemical or physical factory, etc… On the other hand, for all possible POS (part-of-speech), WordNet provides a synset and the corresponding definition for each sense (meaning) of “plant”. In this case, there are four senses for the noun and six senses for the verb “plant”.

![Figure 1. Information for mention “plant” in Wikipedia versus WordNet.](Image)

We summarize below special situations that could be met in employing WordNet and the corresponding rules to be applied. Let the word to be examined be \( w \), and we wish to find its synset.

- If \( w \) is not found in WordNet, nothing is returned. In other words, \( w \) is skipped.
- If \( w \) is found in WordNet, and there is a number of possible senses for it (\( w \) is polysemous): the sense with the largest score produced from the word sense
disambiguation (WSD) process is chosen. The synset corresponding to this sense is returned.
• If \( w \) is found in WordNet and there is only a single sense for it (\( w \) is monosemous): there is no WSD processing needed and the synset corresponding to this sense is returned.

3.3 Synset expansion approach

Our proposed approach concentrates on using WordNet to modify the construction of the bag of words in the statistics stage of the WIN algorithm.

The intuition behind this approach is that due to linguistic matters, authors of various documents might choose to use different versions – or synonyms – of a word sense in the process of creating the documents. Consequently, the approach of using only observed surface forms in a body of text can lead to mismatches when we perform similarity computation. For instance, surface forms “airplane” and “aeroplane” both indicate the flying vehicle and should indeed be considered the same in the matching process.

From this observation, we propose an approach to enrich the provided text features for the mention. Consider the text features representing a mention (previously described in Section 3.1). Instead of only passing that text through a normalizing process (in particular, removing stop words, punctuation, special characters and stemming) to build a bag, we construct a new version of the bag using specific WordNet features. The type of transformation examined in our approach is expansion.

In this expansion approach, for each word in the bag of words, its set of synonyms (its synset) needs to be determined and then added as an expansion to the original bag. In particular, by using synset expansion, multiple surface forms that refer to the same sense can have a higher chance to match in the ranking process.

Due to the following two reasons, only the bag of words for the mention is considered in the expanding process. The first reason is when a word is expanded into its possible synset of equivalent surface forms, it is sufficient to facilitate matching with those forms in the other side, namely, Wikipedia text, if they actually exist there. Consequently, it is not necessary for the expansion of both the bag of words representing the mention and the bag of words representing Wikipedia article.

The second reason is the processing required to extend on the Wikipedia side. In fact, we have to consider the body of Wikipedia article in order to obtain proper context for the outgoing links. Although this processing can be performed offline for Wikipedia articles, in our experiments, we observe that this approach offers no positive result to the system precision. We therefore only concentrate on the approach of expanding the bags of words for mentions in the text.

Our approach consists of two methods to explore on the basic idea of expanding the bag of words for a mention in the text. The first method performs synset expansion on all words in the bag and the second method performs synset expansion only on single-sense words.

In subsequent descriptions of the two proposed methods, assume we are examining a mention \( m \) and its corresponding list of \( k \) Wikipedia candidate articles \( T_i, i = 1..k \). The bag of words representing text features for \( m \) is \( bw(m) \) and the bags of words representing Wikipedia features for Wikipedia articles are \( bw(T_i), i = 1..k \).

Also note that there is a threshold defined to determine whether a similarity value is sufficient for the examined candidate to be the disambiguation result for \( m \). The rules that handle special cases of determining word senses, as stated in Section 3.2, are applied. For example, a word is skipped when its sense cannot be found in WordNet.

In the first method, the following steps are performed:

1. For each word in the text features representing \( m \), find its WordNet sense ID. Find the corresponding synset for this sense ID and replace the original word by this synset.
2. Construct the expanded version of the bag of words \( bw(m) \) using the synsets found above.
3. Perform cosine similarity computations between \( bw(m) \) and each of \( bw(T_i), i = 1..k \) to find the candidate with highest similarity to \( m \). If this maximum is larger than the threshold, return this candidate as the mapping result for \( m \).

For instance, take a simple case where the text contains “…he lifted the band…” The bag of words constructed is “.. lift band …”. Then the
WSD algorithm determines that in this context, "lift" points to the 1st sense of the verb "lift" and its corresponding synset consists of ("raise", "lift", "elevate"). Also "band" points to the 4th sense of the noun "band" and its corresponding synset is ("band", "banding", "stripe"). In this case the expanded version of the bag of words is "...lift raise elevate band banding stripe...".

The second method presents a change made to the first step in the first method. This time we only find the synsets for words that have exactly one sense in WordNet. As previously discussed in Section 3.2, because such words have only a single sense, no WSD processing is required to determine its sense ID in order to retrieve the correct synset. As no WSD is involved, noises introduced from the WSD process are not included in the text features. The synsets for these single sense words are then used to expand the bag of words for the mention in the text.

After we have the expanded version of the bag of words, the rest of the process is similar to the original cosine similarity computation. The Wikipedia candidate with Wikipedia features that achieves highest similarity score to text features of the mention is chosen as the linking result.

We employ the tf×idf scheme to assign weights to terms in a bag of words. For the case of the mention in the text, the weighting scheme is altered to accommodate the adding of synonyms. In particular, the length of the computed vector representing the bag of words should not be affected by the added synonyms. This is because these synonyms share the same meaning with their original words in the text and therefore should not contribute to the length.

4 Experiment Results

All experiments are performed on the D2 dataset of text documents in [2], which contain about 2,000 named entities. Obtained data are from various sources, especially scientific websites or news websites. The dataset was also employed by [2] and thus is used in order to evaluate our performance over the authors’ work.

The input to the disambiguation process is a text with already recognized named entities and the output is a corresponding Wikipedia article for each named entity or NIL if there is no such article.

The WSD program in use is UKB, which utilizes a graph-based word sense disambiguation algorithm [12]. UKB applies the so-called Personalized PageRank on a Lexical Knowledge Base (LKB) to rank the vertices of the LKB and thus perform disambiguation. This algorithm requires as input a POS-tagged version of the text in order to find the senses of words. In this paper, we use the Stanford POS-tagger provided from the Stanford NLP group, and explained in [14].

The measure in which the accuracy, or performance, of the system is presented is the MAA (Micro Averaged Accuracy) measure. Let \( T_{all} \) be the number of gold mappings on a dataset (the number of correct mappings that are done manually), \( T_i \) be the number of incorrect mappings and \( T_c \) be the number of correct mappings that are done by a disambiguation system. Note that \( T_{all} = T_c + T_i \). Definition of the measure is provided in the following formula:

\[
MAA = \frac{T_c}{T_c + T_i} = \frac{T_c}{T_{all}}
\]

Table 1 below summarizes the MAA results of the methods experimented. E1 and E2 indicates the first method (synset expansion of all words) and second method (synset expansion of only single-sense words), respectively. The ALL, PER, LOC and ORG columns contain MAA result for all named entities and for each of the three categories of named entities: person, location and organization.

<table>
<thead>
<tr>
<th>Method</th>
<th>ALL</th>
<th>PER</th>
<th>LOC</th>
<th>ORG</th>
</tr>
</thead>
<tbody>
<tr>
<td>WIN</td>
<td>90.1</td>
<td>92.4</td>
<td>86.8</td>
<td>91.4</td>
</tr>
<tr>
<td>E1</td>
<td>91.6</td>
<td>93.3</td>
<td>88.1</td>
<td>93.2</td>
</tr>
<tr>
<td>E2</td>
<td>92.5</td>
<td>93.9</td>
<td>89.3</td>
<td>94.8</td>
</tr>
</tbody>
</table>

The results show that our two proposed methods utilizing the expansion approach are able to improve the precision of the named entity disambiguation approach from [2]. In the second method, a more special criterion is employed, where only words with a single sense in WordNet are considered in the expansion process. This case may seem to be limited, but it is motivated by the fact that synonymous words added as expansion into the bag of words in this manner have the lowest change of being incorrect. In other words, only relevant information may be added and thus can serve as a more appropriate source.
of enriching data.

In the first method, all words in the text features are considered in the expansion process, which may lead to irrelevant information because of the WSD system’s inability to correctly identify word senses. These wrongly identified word senses are then used to expand incorrect synonyms into the bag of words and the ranking of Wikipedia candidate articles is also affected. Taking these results and analyses into consideration, a WSD algorithm that can offer higher accuracy can help to further improve the performance of the disambiguation process.

We examine the following example case of disambiguating the mention “AP” from an excerpt of a test document (shown in Figure 2). In this case, the original WIN algorithm maps the mention to NIL due to an insufficient amount of text features to use in similarity computation.

![Figure 2. Excerpt from a news document containing the mention “AP”](image)

Our first method uses synset expansion of all words in the text features for the mention. Although the text features are enriched, the mapping result is incorrectly determined as “Andhra Pradesh”, because irrelevant synsets retrieved and added affect the similarity computation and consequently lead to the wrong result. Meanwhile, our second method relies on synset expansion of only single-sense words (e.g. “reporter” expanded into “reporter”, “newsman”, “newsperson”) and helps to increase the similarity with the text features representing the correct Wikipedia article, namely “Associated Press”.

5 Conclusion

We have explored and experimented different ways to exploit WordNet features for named entity disambiguation. The conducted experiments show that our proposed expansion method, utilizing synsets to enrich the text features for the mention, improves the performance of the system in terms of the MAA measure. The strategy that uses only synsets of monosemous words achieves higher performance by avoiding adding irrelevant information to the bag of words model.

In addition to named entities, we are currently investigating methods to use WordNet features for efficient abstract entity disambiguation. The task of disambiguating abstract entities, however, is challenging due to the complex and overlapping nature of the corresponding set of Wikipedia articles.

References


