Unsupervised Word Sense Disambiguation and Rules Extraction using non-aligned bilingual corpus

Francisco OLIVEIRA¹, Fai WONG¹, Yiping LI¹, Jie ZHENG²

¹Faculty of Science and Technology, University of Macau, Macau
²Department of Automation, Tsinghua University, Beijing

sefrancisco@inesc-macau.org.mo, {derekfw, ypl}@umac.mo, zhj@mail.au.tsinghua.edu.cn

Abstract—This paper presents a statistical Word Sense Disambiguation with application in Portuguese-Chinese Machine Translation systems. Due to the limited availability of Portuguese-Chinese resources in the form of digital corpora and annotated Treebank, an unsupervised learning and a non-aligned bilingual corpus are applied. The proposed method first identifies words related to each of the ambiguous words based on their surrounding words and relative distance. A mathematical model is then applied in the identification of the most suitable sense of an ambiguous word in terms of the related words. All the senses discovered are converted into a set of rules and stored in the Sense Knowledge base for later use in disambiguation and translation process. Preliminary experiment results show an improvement of 6% in assigning correctly the corresponding translation over the baseline method.

Index Terms—Word Sense Disambiguation, Natural Language Processing, Machine Translation

I. INTRODUCTION

Many words in natural language are known to be highly ambiguous. For example, the noun "Português" (Portuguese) can either be in the sense of a Portuguese language or in the sense of a human with a Portuguese nationality.

The objective of Word Sense Disambiguation (WSD) is to identify the correct meaning or sense of a word in a given sentence. In fact, WSD is crucial for Information Retrieval systems that require the correct understanding of a sentence and for Machine Translation systems that require the selection of the most appropriate word to translate in the target language.

Due to the availability of large amount of documents available in electronic format, many of the reported WSD models are corpus-based approaches. These models use machine-learning techniques to automatically acquire such disambiguation knowledge which is classified as supervised or unsupervised learning.

In supervised approach, word senses are identified by using a large manually tagged sense corpus. However, this approach is not practical due to the expensive cost and a time consuming task in manually labeling senses. To get around the problem, Brown et al. [1] and Gale et al. [2] used a bilingual parallel corpus aligned in sentence and token level.

In unsupervised approach, the disambiguation is done without the use of a sense tagged corpus. Most of them rely on the use of machine-readable dictionaries. Dagan and Itai [3] used a bilingual lexicon and parsers in addition to a bilingual parallel corpus. Kaji and Morimoto [4] also used a bilingual dictionary and their method only requires a bilingual comparable corpora being on the same domain. Kikui [5] and Tanaka and Iwasaki [6] relied on monolingual corpora to tackle the problem.

In terms of the techniques applied, many of them used their mathematical model in identifying the most suitable sense for each ambiguous word in a given context. Gale et al.'s approach is based on a Naive Classifier, while Kaji and Morimoto mapped the correlations between the senses and clues into a mathematical formula. Brown et al. applied a flip-flop algorithm and a splitting theorem [1] for disambiguation binary senses which is similar to a decision tree learning method. Yarowsky [7] relied on decision lists for the resolution of ambiguities. Kikui [8] even combines the approach based on a distributional clustering proposed by Schuetze [9], and based on a score describing relationships between coherent words [5].

The discovery of words correlated ambiguous word is also a vital issue in WSD. Different considerations had been proposed, including the identification of surrounding words, local collocations, and syntactical relations. Dagai and Itai, Kaji and Morimoto used a context window to find out their relationships. Ng and Lee [10] made use of local collocations to disambiguate word sense. Brown et al. used a part of speech tagger in order to identify the relationships between the related and the ambiguous word. Kikui mapped words into a multidimensional vector space in order to find out the coherence in terms of geometrical relationships between the words. However, there are few literatures discussing about the combination of some of the mentioned techniques.

Due to the limited availability of aligned Portuguese and Chinese bilingual resources in the form of digital corpora, and an annotated Treebank, an unsupervised approach is applied using a set of bilingual sources, including a non-aligned bilingual corpus, a bilingual dictionary, and a sense inventory. Moreover, in order to obtain better results in the identification of related words, this paper proposes an approach that considers not only the surrounding words of the ambiguous word within a
context window, but also their relative distance to tackle syntactical relationships.

This paper is organized as follows: section II describes the disambiguation algorithm based on a mathematical formula. Section III presents the core design of the proposed approach. Section IV gives the evaluation and experimental results, and possible future improvements. An application of the proposed method will be given in section V, and finally, there will be a conclusion.

II. DISAMBIGUATION ALGORITHM

The disambiguation process is based on a mathematical formula that calculates the relationship between the ambiguous and its related word.

An ambiguous word is defined as the one having more than one meaning given by a bilingual dictionary. Related words are considered as those having a close relationship with the ambiguous words. They are considered crucial in finding out the sense that best fits it. For example, in the sentence "Hoje eu tenho tempo livre" (Today I have free time), the ambiguous word Wi" tempo" can be translated as either "天氣" (weather) or "時間" (time). If the related word(s) Wi of "tempo" can be identified, then it provides clues for selecting the correct sense. In this case, if Wi is "livre" (free), then it is a good clue for selecting the translation of "tempo" as "時間". Given the pair (tempo, livre), our model lookups in the bilingual lexicon for all the possible translations. Suppose that their translations are: "時間" (time), "天氣" (weather) and "自由的" (free), "空閒的" (free). Thus, there are totally 4 translation alternatives: (時間, 自由的), (時間, 空閒的), (天氣, 自由的), (天氣, 空閒的). A score is then assigned for each of these to determine which one can best resolve the sense of the ambiguous word.

Given a sentence S, if there is an ambiguous word Wi, then there should be a related word W, which can resolve the ambiguity of Wi. For each Wi, there is a set of different senses C={c_i, ..., c_a}. We denote P(c_i | Wi) as the score of one possible sense c_i given Wi. Thus, the best sense c_i can be computed as

\[ c_i = \arg \max_{c_i} P(c_i | W_i) \tag{1} \]

If Wi has only one sense and every c can be obtained by a different W, we can approximate P(c_i | W_i) in terms of frequency. Let freq(d) be the number of times that word d has appeared in the corpus, as shown below:

\[ P(c_i | W_i) = \frac{\text{freq}(c_i, W_i)}{\text{freq}(W_i)} = \frac{\text{freq}(c_i, c_i)}{\text{freq}(c_i)} \tag{2} \]

freq(c_i, c_i) and freq(c_i) denote the number of times that c_i, c_i, and c_i have appeared in the corpus. Since freq(c_i) is a constant, we can rewrite the equation as

\[ c_i = \arg \max_{c_i} \text{freq}(c_i, c_i) \tag{3} \]

If Wi has more than one sense and each c can be obtained by any W, we have to transform (3) into a two-dimensional equation:

\[ (c_i, c_j) = \arg \max_{(c_i, c_j)} P(c_i, c_j | W_i, W_j) \]

\[ = \arg \max_{(c_i, c_j)} \frac{P(c_i, c_j, W_i, W_j)}{P(W_i, W_j)} \tag{4} \]

\[ = \arg \max_{(c_i, c_j)} P(c_i, c_j, W_i, W_j) \]

If each pair (c_i, c_j) can be obtained by any pair (W_i, W_j), W_i and W_j can be a word being sense c_i and c_j, then we can approximate P(c_i, c_j, W_i, W_j) as:

\[ P(c_i, c_j, W_i, W_j) = P(c_i, c_j) \times \frac{P(W_i, W_j)}{\sum_{W'_i, W'_j} P(W'_i, W'_j)} \tag{5} \]

If we take into consideration of the number of times appeared by (W_i, W_j) in the training corpus, we can define the score of using the senses c_i and c_j for words W_i and W_j as:

\[ (c_i, c_j) = \arg \max_{(c_i, c_j)} \text{freq}(c_i, c_j, W_i, W_j) \]

\[ = \arg \max_{(c_i, c_j)} \frac{\text{freq}(c_i, c_j, W_i, W_j)}{\sum_{W'_i, W'_j} \text{freq}(W'_i, W'_j)} \tag{6} \]

Based on the value calculated for each translation candidate pair, the one that has the highest score will be selected. For example, since the related word "livre" (free) with meaning "空閒的" can best disambiguate the word "tempo" (time) with meaning "時間", it should have the highest value among the other candidates.

III. PROPOSED FRAMEWORK

The whole process is divided into two phases: acquisition and translation phase, as shown in Fig. 1. During the first stage, it first extracts all the ambiguous words from the source corpus. For each of these words, their corresponding related words are identified based on a defined context window and their relative distance. The translations and the senses of these words are extracted from a bilingual dictionary and a sense inventory.
Disambiguation of the ambiguous word is done by selecting the combination of the translation alternative that has the highest score defined by a mathematical formula. Given a sense, its related semantic and syntactical information are converted into a rule format and stored in the Sense Knowledge base. During the translation stage, based on the word given, it searches all the rules stored in the WSD database and retrieves the translated word.

A. **Corpus Pre-processing**

During the pre-processing phase, it first divides the whole bilingual text into fragments of texts according to the punctuation (full stop, semicolon, etc). Once identified, the process removes non-informative words by comparing with a list of “stop words”. For each word in a Portuguese sentence, their original form will be identified by applying a morphological analysis. This helps in improving the exact matching between words in the source corpus and bilingual dictionary. As an example, the sentence “Muitos computadores” (Many computers) becomes “Muito computador” after morphological analysis. Finally, a Tagger is applied to find out their corresponding part of speech.

B. **Ambiguous and Related words identification**

Ambiguous words are identified by using a bilingual dictionary. Two considerations are taken in the selection of the most related word: a context window, i.e. a list of words around the ambiguous word within a boundary; the degree of relatedness by considering their relative distance. Consider the following morphologized Portuguese sentence:

\[
\text{[Se ter tempo livre, ir] à praia.} \\
\text{(If have free time, go to the beach.)}
\]

Here, a context window (size=2, indicated by square brackets) is defined to find out the set of related words surrounding the ambiguous word. Its related words are the following: “se” (if), “ter” (to have), “livre” (free), and “ir” (to go). Moreover, priority is assigned to these related words. This is due to the following assumption: the closer is the relative distance between the candidate word and the ambiguous word, the closer is their relation. In other words, those that are closer to the ambiguous word will be first given to the disambiguation module. In this example, the order is the following: “ter”, “livre”, “se”, and “ir”. If the closest related word cannot resolve the sense of the ambiguous word, the next closest related word will be chosen. This iteration only terminates if the candidate word helps in resolving the sense or there are no words related to the ambiguous word. Using the above approach, importance is given not only to the use of a context window, but also to the words that are closely related in terms of distance. It also helps the algorithm to have better achievements in choosing the correct sense because it also considers the syntactical relationships between words in a sentence.

C. **Sense categorization**

For each ambiguous and related word found, their corresponding senses will be extracted from the sense inventory. This inventory is developed by the University of Macau and “Instituto de Engenharia de Sistemas e Computadores de Macau” (INESC-Macau) Speech and Language Learning team. From the viewpoint of Machine Translation Systems, each Portuguese word is assigned with a sense based on its part of speech and meaning. Totally, there are more than 50 different senses defined in the sense inventory, where some of them are shown in Table I.
Table 1

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Sense</th>
<th>Abbreviation</th>
<th>Sense</th>
</tr>
</thead>
<tbody>
<tr>
<td>b'</td>
<td>Time</td>
<td>d'</td>
<td>Culture</td>
</tr>
<tr>
<td>d</td>
<td>Place</td>
<td>h</td>
<td>Animal</td>
</tr>
<tr>
<td>n</td>
<td>Nature</td>
<td>p</td>
<td>Money</td>
</tr>
<tr>
<td>c</td>
<td>Emotion</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Suppose the entries found in the bilingual dictionary of the word “Portugués” (Portuguese) are: “葡萄牙人” (Portuguese people), “葡萄牙語” (Portuguese language), and “葡萄牙的舊金幣” (Old coin of Portuguese). Consequently, in the sense inventory, it contains the following senses for the word “Portugués”: “h” (human), “w” (language), and “$” (money) respectively.

D. Scoring each translation candidate

Base on the identified related word, the model searches in the bilingual lexicon and the sense inventory for all the possible translations and senses of the ambiguous and related word. For each of these alternatives, a score is assigned using (6) and the training corpus in order to determine the one that can best resolve the sense of the ambiguous word.

Once the best translation pair is retrieved, the corresponding senses of the ambiguous and related words are selected from the sense inventory.

E. Rules generation

All the generated information in the previous stage is then stored in the Sense Knowledge base. Since this information can affect the accuracy of the translation process, a careful design of the database is necessary.

According to Ide and Veronis [11], there are three important characteristics of an ambiguous word: grammatical information about the ambiguous word to be disambiguated, words that are syntactically related, and words that are topically related to the ambiguous word. Since the proposed method relies on the semantic and syntactical information to disambiguate an ambiguous word, consideration is taken in the first two types. For each entry of the Sense Knowledge base, it consists of the following: ambiguous and related word, sense of the ambiguous and related word, and part of speech of the related word. Moreover, this information can be converted into an understandable rule format that best describes the relationship between the ambiguous word and the related word. As an example, consider the following rule of the ambiguous word “tempo” (weather):

SELECT (tempo/N/天氣, n)
IF (W1="tempo" AND Wr="bom") OR (POS(Rel)=Adj AND Sense(Rel)=Sense("bom"))

Based on the information retrieved from an entry of the Sense Knowledge base, rules can be easily constructed to describe the relationship between the ambiguous and related word. Here, sense “n” (nature) is assigned to the ambiguous word “tempo” if either of the following cases is true: the most related word is “bom” (good); the part of speech and sense of the proposed related word are the same as “bom”. The later condition is used to relax the first constraint. For example, even if the most related word identified is not “bom”, it can still find out the corresponding sense if the selected related word is an adjective and has the same sense as “bom”. In this case, even if the related word is “mal” (bad) or “chuvoso” (rainy) or “nublado” (cloudy), the sense of “tempo” can still be identified.

F. Translation phase

The translation process makes use of the Sense Knowledge base constructed in the previous phase and a bilingual lexicon to disambiguate all the ambiguous words found in a given sentence. As an example, consider the following sentence:

“Eu falo portugues” (I speak Portuguese)

Suppose that the ambiguous word is “portugues” (portuguese). The process first performs a part of speech and morphological analysis to assign the correct part of speech and to restore the original format of the sentence. This is a useful procedure to eliminate the ambiguity of the words. In this example, since the word “portugues” is tagged as noun, all the meanings related to the word “portugues” treated as adjective will be discarded. Next, it searches for all the entries in the Sense Knowledge base if there are any ambiguous and related words associated to “portugues”. If there is an exact match for the ambiguous word “portugues” and related word “falar”, it returns “我會講葡萄牙語” (I speak portuguese).

IV. Experiment Results

Experiments are done by using a bilingual training corpus related to sentences extracted from a Portuguese-Chinese Grammar book. Totally, 1900 sentences are extracted and 100 sentences are randomly selected and considered as testing data. The windows size applied in the experiments is 3. The choice of the windows size is not a big value due to the consideration of each ambiguous word is made in a sentence level rather than a whole document level. Moreover, if the windows size is becoming larger, more unusable information will be generated.

A. Number of rules extracted

During the acquisition process, we found that as the number of the sentences used in the training increases, the number of new rules generated tends to decrease, as shown in Fig. 2. This is mainly due to the domain and the inherent similarity of the sentences used in the training.
B. Evaluation of the proposed method

Two measurements are applied in order to evaluate our method: applicability and precision [3]. The applicability is defined as the number of cases that the algorithm could disambiguate. The precision is the proportion of previously found cases that the algorithm disambiguated correctly.

A comparison with the baseline method is also performed. Baseline method always assigns the most frequent sense to each of the ambiguous words.

The performance of the baseline and the proposed method is summarized in Tables II and III. The applicability of the baseline method is 100%, while our method only achieves 76.9%. However, in terms of precision, our method is about 6% better than the baseline method.

C. Observations

The proposed method has some limitations. One of these is the assumption we considered previously: the closer is the relative distance between the candidate word and the ambiguous word, the higher is their relation. Although it helps to tackle the syntactical relationships between the related and ambiguous word, it sometimes may fail to find out the most related word. For example, consider the following sentence: “tenho tempo livre” (have free time). If the ambiguous word is “tempo”, the system may probably treat the verb “ter” has a higher relationship than the word “livre”. This can be solved by considering their mutual information.

Another issue is related to the definition of the context window. For example, consider the sentence shown before:

>[Se ter tempo livre, ir] à praia.

Since it consists of two sub-sentences, for the ambiguous word “tempo”, the proposed method shouldn’t treat the word “ir” (to go) as one possible candidate related word. One possible solution is to make use of a parser to further identify the internal relationships of the sentence.

V. APPLICATION OF THE PROPOSED METHOD

In Macau, there is a Portuguese to Chinese Machine aided translation system called PCT Assistente. It is a system that provides professional translators a workbench for their translation work. Moreover, it applies sophisticated technologies, like a morphological analyser, the use of a Translation Correspondence Tree [12], a Constraint-based Synchronous grammar, etc. The proposed method can be used to further enhance the translation quality of the system.

VI. CONCLUSION

In this paper, a framework for choosing the most suitable sense and translation of a word based on an unsupervised learning using non aligned corpus is presented. A context window and the relative distance between the related and ambiguous words are considered in order to tackle the syntactical relationships and dependencies between them within the context defined. The selection of the correct senses from a pair of related words is done by using a mathematical formula and a set of bilingual sources, including a lexicon, a sense inventory, and a training corpus. Moreover, this framework is applicable and useful for enhancing the translation quality in PCT Assistente, a Portuguese-Chinese Machine Translation System. Experimental results indicate that the proposed method improves the precision compared with the baseline method.

REFERENCES


Authorized licensed use limited to: Universidade de Macau. Downloaded on June 19,2010 at 13:24:32 UTC from IEEE Xplore. Restrictions apply.


