Proceedings of the Fourth International Conference on Machine Learning and Cybernetics, Guangzhou, 18-21 August 2005

AN UNSUPERVISED & STATISTICAL WORD SENSE TAGGING USING BILINGUAL SOURCES

FRANCISCO OLIVEIRA, FAI WONG, YI-PING LI

Faculty of Science and Technology, University of Macau, Macau SAR
E-MAIL: sefrancisco@inesc-macau.org.mo, derekfw@umac.mo, ypli@umac.mo

Abstract:

This paper presents an approach for choosing the correct translation of an ambiguous word in a given sentence. An unsupervised learning is applied and a non-aligned bilingual Portuguese to Chinese bilingual corpus is used in disambiguating word senses. The identification of the relationships between words is done by considering its surrounding words and their relative distance to tackle syntactical relationships. All the related words are then translated to the target language in finding out the correct senses of ambiguous words. The selection is based on a statistical and a mathematical model by assigning a score to each of the sense identified previously. After all the senses discovered, its semantic and syntactical information are converted into a set of rules and stored in the database for later use in the disambiguation process. Preliminary experiment results of the proposed method shows an improvement of 6% in assigning correctly the corresponding translation over the baseline method.

Keywords:

Word Sense Tagging; Machine Translation

1. Introduction

Word Sense tagging can be viewed as a special case of Word Sense Disambiguation (WSD) that is considered as a difficult task in Natural Language Processing. In fact, the choice for selecting the most appropriate word to translate in the target language is crucial for Machine Translation Systems.

The nature of statistical Word Sense Tagging depends on the leaning techniques applied in the acquisition phase, which is broadly classified as supervised or unsupervised learning. In the first type of learning, word senses are identified by using a large manually tagged sense for each ambiguous word. For example, Brown et al. [1] and Gale et al. [2] used a bilingual parallel corpus in disambiguating word senses. However, this approach is not practical due to the expensive cost and a time consuming task in manually labeling senses. On the other hand, in unsupervised learning, the disambiguation is done without the use of a sense tagged corpus. Dagan and Itai [3] used a bilingual lexicon and parsers in addition to a bilingual parallel corpus. Kaji and Morimoto [4] used the bilingual comparable corpora that only require being on the same domain and a bilingual dictionary. 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. Gale et al.'s approach is based on a Naïve 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] 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].

Another vital issue in statistical Word Sense Tagging relies on the techniques applied in identifying words correlated to the ambiguous word. Brown et al. used a part of speech tagger in order to identify the relationships between the related and the ambiguous word, while Dagai and Itai, Kaji and Morimoto used a context window to find out their relationships. 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.

Moreover, many of the previous reports were targeted in using either a monolingual corpus or a bilingual corpus in English, German, Spanish, French, Hebrew, Chinese, Japanese or combinations between these languages. And very few of them are related to the use of a Portuguese and Chinese language pair due to the structural (character types) and syntactical differences and the limited availability of resources in the form of digital corpora, computational lexicons, grammars or annotated Treebank.

0-7803-9091-1/05/$20.00 ©2005 IEEE

This paper describes an unsupervised Word Sense Tagging by using a set of Portuguese-Chinese bilingual sources: a training corpus, a dictionary, and a sense inventory. The whole process is divided into two phases: acquisition and tagging phase. 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. Once the related word is found, the translations of these words are extracted from a bilingual dictionary. 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 WSD database. During the tagging stage, based on the word given, it searches all the rules stored in the WSD database and retrieves the corresponding translated word.

This paper is organized as follows: section 2 describes the senses defined in the sense inventory. Section 3 presents the core and detailed design of the proposed approach. Section 4 gives the evaluation and experimental results, and possible future improvements. Finally, there will be a conclusion.

2. Senses Categorization

From the viewpoint of Machine Translation Systems, each Portuguese word is assigned with a sense based on its part of speech and meaning. All these entries are stored in a sense inventory containing more than 50 different senses, as shown in Table 1.

Table 1. Part of the senses defined in the sense inventory

Abbreviation	Sense	Abbreviation	Sense
b	Department	t	Time
d	Place	u	Culture
h	Human	x	Animal
n	Nature	w	Language
p	Plant	$	Money
c1	Emotion	b1	Medical

Suppose the entries found in the bilingual dictionary of the word "português" (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 "português": h, w, and $ respectively.

3. Proposed framework

We define an ambiguous word if it has more than one meaning given by a bilingual dictionary. The identification of the word related to the ambiguous word is crucial in finding out the sense that best fits it. For example, consider the following sentence:

"O *tempo* está bem e temos *tempo* livre"
(The weather is good and we have free time)

Here, the ambiguous word "tempo" can be translated as either "天氣" (weather) or "時間" (time). If the related word(s) of "tempo" can be identified, then it provides clues for selecting the correct sense. In this case, the pair (tempo, bem) is a good clue for selecting the translation as "天氣", while (tempo, livre) is a good clue for selecting "時間".

The whole process of the proposed method is shown in Figure 1.

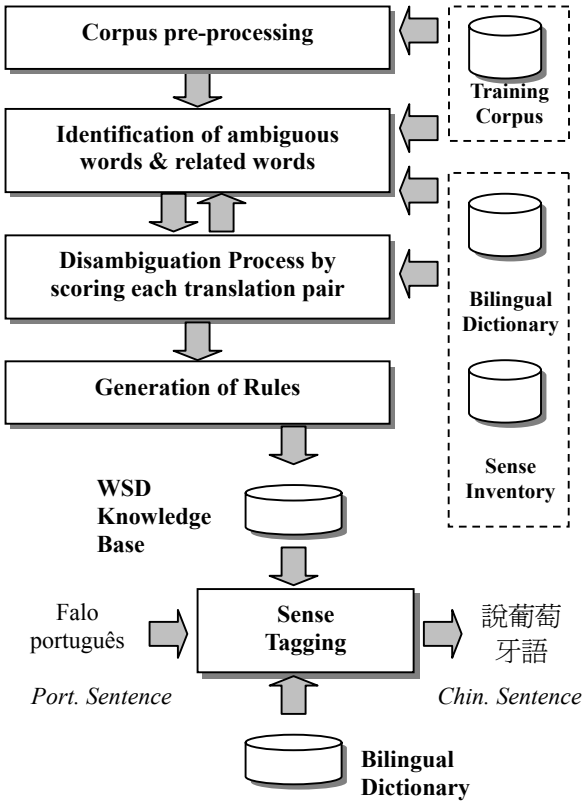

Figure 1. Overall process of the proposed method

3.1. Preliminary processing

During the preprocessing 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 following sentence "Muitos computadores" (Many computers) becomes "Muito computador" after morphological analysis. Finally, a Tagger is applied to find out their corresponding part of speech.

3.2. Identification of ambiguous and related words

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 morpholoolized Portuguese sentence:

[*Se ter tempo livre, ir***]** *à praia.*
(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.

3.3. Translation alternatives

Given a related word (W_r) and the ambiguous word (W_i), our model lookups in the bilingual lexicon for all the possible translations of both words. Suppose that W_r is "livre" (free) and W_i is "tempo" (time), and its translations are: "時間" (time), "天氣" (weather) and "自由的" (free), "空閒的" (free). Totally, there are 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.

3.4. Scoring each translation candidate

The score of each translation candidate is defined by a mathematical formula. Given a sentence *S*, if there is an ambiguous word W_i, then there should be a related word W_r that can resolve the ambiguity of W_i.

For each ambiguous word W_i, there is a set of different senses $C=\{ c'_{i1}, \ldots, c'_{it}\}$. We denote $P(c'_i \mid W_r)$ as the score of one possible sense c'_i given W_r. Thus, the best sense c_i can be computed as

$$c_i = \arg\max_{c'_i} P(c'_i \mid W'_r) \quad (1)$$

If W_r has only one sense and every c can be obtained by a different W, we can approximate $P(c'_i \mid W_r)$ in terms of frequency. Let *freq(d)* be the number of times that word *d* has appeared in the corpus, as shown below:

$$P(c'_i \mid W_r) = P(c'_i \mid c'_r) = \frac{P(c'_i, c_r)}{P(c_r)} = \frac{freq(c'_i c_r)}{freq(c_r)} \quad (2)$$

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

$$c_i = \arg\max_{c'_i} freq(c'_i c_r) \quad (3)$$

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

$$\begin{aligned}(c_i, c_r) &= \arg\max_{(c'_i, c'_r)} P(c'_i, c'_r \mid W_i, W_r) \\ &= \arg\max_{(c'_i, c'_r)} \frac{P(c'_i, c'_r, W_i, W_r)}{P(W_i, W_r)} \\ &= \arg\max_{(c'_i, c'_r)} P(c'_i, c'_r, W_i, W_r)\end{aligned} \quad (4)$$

If each pair (c'_i, c'_r) can be obtained by any pair (W'_r, W'_i), W'_i and W'_r can be a word being sense c'_i and c'_r Thus we can approximate $P(c'_i, c'_r, W'_r, W'_i)$ as:

$$P(c'_i, c'_r, W_i, W_r) = P(c'_i, c'_r) \times \frac{P(W_i, W_r)}{\sum P(W'_i, W'_r)} \quad (5)$$

If we take into consideration of the number of times appeared by (W'_i, W'_r) in the training corpus, we can define the score of using the senses c'_i and c'_r for words W_i and W_r as:

$$(c_i, c_r) = \arg\max_{(c'_i, c'_r)} P(c'_i, c'_r, W_i, W_r)$$
$$= \arg\max_{(c'_i, c'_r)} \frac{freq(c'_i\, c'_r) \times freq(W_i W_r)}{\sum freq(W'_i\, W'_r)} \quad (6)$$

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. Once the best translation pair is retrieved, the corresponding sense of the word "tempo" is selected from the sense inventory.

3.5. Rules Generation

All the generated information in the previous stage is then stored in the WSD 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 [10], 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 WSD 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 (Wi="tempo" *AND* Wr="bom") OR (POS(Rel)=Adj *AND* Sense(Rel)=Sense("bom"))

Based on the information retrieved from an entry of the WSD 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.

3.6. Translation Process

The translation process makes use of the WSD 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: "Bom tempo" (Good weather). Suppose that the ambiguous word is "tempo" (weather). The process first performs a morphological analysis to restore the original format of the sentence. Next, it searches for all the entries in the WSD Knowledge base if there are any ambiguous and related words associated to "tempo". Since there is an exact match for the ambiguous word "tempo" and related word "bom" (good), it returns "好的天氣" (Good weather).

4. Experimental 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.

4.1. 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 Figure 2. This is mainly due to the domain and the inherent similarity of the sentences used in the training.

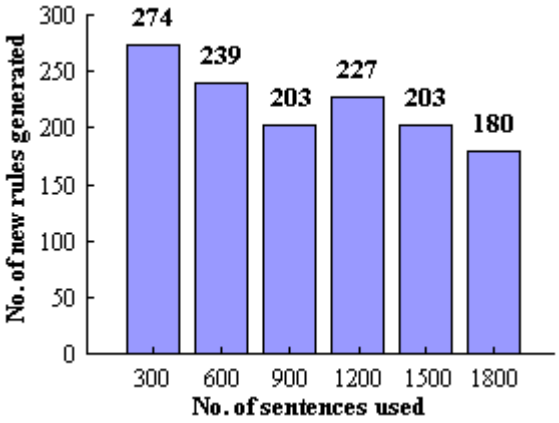

Figure 2. New rules generated in acquisition phase

4.2. 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 2 and 3.

Table 2. Experiment results – Applicability

	Baseline method	Proposed method
No. of ambiguous words	352	352
Disambiguated words	352	271
Applicability	100%	76.9%

Table 3. Experiment results – Precision over the disambiguated words found by the proposed method

	Baseline method	Proposed method
Disambiguated words	271	271
Words disambiguated correctly	205	223
Precision	75.6%	82.3%

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.

4.3. 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 previously: "[Se ter *tempo* livre, ir] à praia." (If we have free time, we go to the beach). Since the sentence 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.

5. Application of the proposed method

In Macau, there is a Portuguese to Chinese Machine aided paragraph 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 [11], a Constraint-based Synchronous grammar, etc. The proposed method can be used to further enhance the translation quality of the system.

6. Conclusion

In this paper, a framework for choosing the correct translation of a word based on an unsupervised learning and the use of bilingual sources 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. Experimental results indicate that the proposed method improves the precision compared with the baseline method.

References

[1] P. Brown, S. D. Pietra, V. D. Pietra, and R. Mercer, "Word Sense Disambiguation Using Statistical Methods", Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, pp. 264-270, 1991.

[2] W. Gale, K. Church, and D. Yarowsky, "A Method for Disambiguating Word Senses in a Large Corpus", Computers and Humanities, vol. 26, pp. 415-439, 1992a.

[3] I. Dagan and A. Itai, "Word Sense Disambiguation using a Second Language Monolingual Corpus", Computational Linguistics, vol. 20, pp. 563-596, 1994.

[4] H. Kaji and Y. Morimoto, "Unsupervised Word Sense Disambiguation Using Bilingual Comparable Corpora", Proceedings of the 19th international conference on Computational Linguistics, pp. 411-417, 2002.

[5] G. Kikui, "Term-list translation using monolingual co-occurrence vectors", Proceedings of the 17th international conference on Computational Linguistics, pp. 670-674, 1998.

[6] K. Tanaka and H. Iwasaki, "Extraction of lexical translations from non-aligned corpora", Proceedings of the 16th international conference on Computational Linguistics, pp. 580-585, 1996.

[7] D. Yarowsky, "Decision Lists for Lexical Ambiguity Resolution: Application to Accent Restoration in Spanish and French", Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics, pp. 88-95, 1994.

[8] G. Kikui, "Resolving Translation Ambiguity Using Non-parallel Bilingual Corpora", Proceedings of ACL'99 Workshop on Unsupervised Learning in Natural Language Processing, 1999.

[9] H. Schutze, "Automatic Word Sense Discrimination", Computational Linguistics, vol. 24, no. 1, pp. 125-146, 1998.

[10] N. Ide, N. and J. Veronis, "Introduction to the special issue on word sense disambiguation: the state of the art", Computational Linguistics, vol. 20, no. 4, pp. 563-596, 1994.

[11] F. Wong, D. C. Hu, Y. H. Mao, and M. C. Dong, "A Flexible Example Annotation Schema: Translation Corresponding Tree Representation", Proceedings of 20th international conference on Computational Linguistics, pp. 1079-1085, 2004.