A Word Sense Disambiguation Method Using Bilingual Corpus

Zheng Jie, Mao Yuhang

Institute for Language and Speech Processing

Department of Automation, Tsinghua University, P.R.China

zhj@mail.au.tsinghua.edu.cn

Abstract

This paper proposes a word sense disambiguation (WSD) method using bilingual corpus in English-Chinese machine translation system. A mathematical model is constructed to disambiguate word in terms of context phrasal collocation. A rules learning algorithm is proposed, and an application algorithm of the learned rules is also provided, which can increase the recall ratio. Finally, an analysis is given by an experiment on the algorithm. Its application gives an increase of 10% in precision.

1. Introduction

The WSD is a semantic concern which has been involved in a wide range of applications, such as natural language understanding, machine translation, speech processing and so on. A variety of approaches have been proposed to address this problem. Generally, methods of WSD can be classified into two categories: rule-based method and statistics-based method. The advantage of Rule-based WSD method is of its high precision and long distance relatedness, but its disadvantage is strenuous manual work, low coverage and interference of rules. Statistics-based method provides a supplement way to make up the disadvantages, which shows a high coverage and detail knowledge granular resolution. So our research on WSD is focused on the combination of the two methods.

Because it needs a lot of manpower to tag word sense of a large corpus in statistics-based WSD method, efforts have been made to automatically sense-tag a training corpus via bootstrapping methods. Hearst proposed an algorithm that includes a training phase during which each occurrence of a set of nouns to be disambiguation is manually sense-tagged in several occurrences. Statistical information extracted from the context of these occurrences is then used to disambiguate other occurrences. Yarowsky proposed transformation-based method, which assumes that the similar sense would appear more likely in similar context whether it belongs to a polysemous word or single-sense word. So the context information of single-sense word can be applied to the disambiguation of polysemous word. Church proposed the use of bilingual corpora (Ide, Nancy. 1998) to avoid hand tagging of training data. His premise is that the different senses of a given word often translate differently in another language (for example, *pen* in English is *stylo* in French for its 'writing implement' sense, and *enclose* for its 'enclosure' sense). By using a parallel-aligned corpus, the translation of each occurrence of a word such as *pen* can be used to automatically determine its sense. Church's method shows high precision. But it also has the disadvantage due to few large-scale parallel corpora available for use. To decline the requirement for a bilingual parallel-aligned corpus, this paper proposes a WSD method used in English-Chinese machine translation, which uses syntactic relation and bilingual materials. The experiment proves that this method is effective. Section 2 of the paper describes the model for the disambiguation; Section 3 renders the algorithm of the WSD; Section 4 gives an experiment result of the model; Conclusion is given in Section 5.

2. Sense Disambiguation Model

Proposed WSD method needs support from following respects: 1) Bilingual corpus; 2) Sense coding system; 3) Statistical model. The following sections give the detail description.

2.1. Sense coding

Currently most researchers in WSD are relying on the sense distinctions provided by established lexical resources, such as machine-readable dictionaries or WordNet. But the problem of sense division is still an object of discussion. Some people suggest that the sense divisions in dictionaries are too fine for the purpose of natural language processing. Overly fine sense distinctions cause practical difficulties for automated WSD. So we define sense categories in terms of the words' part of speech (POS) and their actual meaning in consideration of its application in English-Chinese machine translation system. The catalogue is simple but it is effective in application. All the words' senses are divided into about 70 classes. A machine readable sense dictionary has been created according to this catalogue. Table 1 gives some samples of the code instances. The authors investigate the

Noun		Adjective	
Code	Category	Code	Category
h	Humankind	a	Commendatory
Y	Vehicle	b	Derogatory
e	Material	f	Negative
x	Animal	h	Sentiment
r	Organ	k	Spacious
y	Tools	o	Probability
$	Money	y	Color

Table 1 Samples of Sense Coding

situation of polysemous words of English text according to sense definition of this English-Chinese electronic dictionary. The result shows an average of 36% of all words have sense or POS ambiguity. After having the text processed by POS disambiguation, 20% still remains the sense ambiguity. The task of sense disambiguation is to select the proper sense from the ambiguous word's sense set.

2.2. Sense Selection through Context

Context is the only means to identify the meaning of a polysemous word. Therefore, all work on sense disambiguation relies on the context of target word to provide information for its disambiguation. Broadly speaking, context can be considered as words in a window surrounding the target word, regarded as a group without consideration for their relationships to the target in terms of grammatical relations. But in natural language, there is some certain rules for a word to form phrase or sentence together with other words. A word may correlate with some kinds of words to form a phrase while it can not apply to other kinds of words. Correlative relation in a sentence describes the basic laws that control the linear word arrangement in natural language. Sense-driven glossary theory indicates that the legitimacy of word combination to form a phrase is not only determined by words' POS, but also by their senses. So context is considered in terms of some relation to the target to form a phrasal collocation here.

This idea can be adopted into the application of sense disambiguation. If a polysemous word is taken into account together with other related words in the context, the most reasonable sense can be selected by appraising all the possible sense correlation. For example, the word "drive" in English is "驱赶" in Chinese for its "force animal to go" sense, and "驾驶" for its "control and guide a vehicle" sense. If "drive" appears in context "drive a car", the context corresponds to two possible Chinese translation: "驱赶汽车" and "驾驶汽车". If a method can be used to appraise the two possible results of translation and gives a higher score to translation "驾驶汽车" than that of translation to "驱赶汽车", "drive" should be chosen the sense of "驾驶".

2.3. Mathematical Description of Sense Selection

An English sentence S can be denoted in mathematical form as: $S=W_1 \ldots W_i \ldots W_N$, where W_i denotes the ith word in the sentence. Each W_i has its POS set P_i and its sense set C_i, they take value from finite discrete symbol set: $P_i=\{p_{i,1}, p_{i,2}, \ldots, p_{i,m_i}\}$, $C_i=\{c_{i,1}, c_{i,2}, \ldots, c_{i,m_i}\}$, where $p_{i,j}$ represents a POS tag of the word W_i, and $c_{i,j}$ represents one of its sense tags. The sentence's correlation R is defined as:

$$R=(p_1', c_1', \ldots, p_i', c_i', \ldots, p_N', c_N')$$

In which p_i', c_i' denotes the POS value and sense value of word W_i in the sentence, $p_i' \in P_i$, $c_i' \in C_i$, $i \in [1, N]$. With these definition, the problem of word sense disambiguation can be described in the following manner: To a given sentence S, find the most reasonable correlation R_m of S, $R_m=(p_{1m}, c_{1m}, \ldots, p_{im}, c_{im}, \ldots, p_{Nm}, c_{Nm})$.

Let $\Phi(R/S)$ denotes the probability of correlation R of the given the sentence S. Then, R_m can be rewritten as R_m. Expand it into POS and sense expressions respectively:

$$p_{im} = \arg\max_{p_i'} \Phi(p_i' \mid S) \quad (1)$$

$$c_{im} = \arg\max_{c_i'} \Phi(c_i' \mid S) \quad (2)$$

Equation 1 gives the model of POS disambiguation. Many methods have been discussed about this problem either by rule-based or by statistics-based model. This paper will only discuss the problem of WSD after POS of words in the sentence have already been determined. Equation 2 describes WSD probability model. From Equation 2 it can be inferred that the sense of word W_i is determined by the sentence to which W_i belongs. Continue to expand Equation 2:

$$c_{im} = \arg\max_{c_i'} \Phi(c_i' \mid W_1 W_2 \ldots W_N) \quad (3)$$

2.4. Context constraints

In fact, the sense selection of word W_i has close relation with some words in the sentence while has little relation with others. From the section 2.2, the sense of W_i can be determined by its context. So it is necessary to simplify the problem by determining related word of W_i. Because the phrasal collocation is stable among POS, it is possible to find the closest related word with W_i by using shallow or partial parsing. To a given noun, for example, its closest related word can be determined by the following phrasal collocation:

Noun positioned leftward	Noun positioned rightward
~ +verb	verb+~
~ +noun	adjective+~
	noun+~
	preposition+~

The symbol '~' denotes the current noun. Through context analysis, choose the closest related word with W_i to be its closest related word, which we call relation word of W_i, and denoted by W_r, the expression 3 can be approximated:

$$c_{im} = \arg\max_{c_i'} \Phi(c_i' \mid W_r) \quad (4)$$

The above equation indicates: By context constraints, the sense of W_i can be determined by its relation word W_r.

2.5. Statistical Disambiguation Model

Probability $\Phi(c_i' \mid W_r)$ can be estimated through the statistical analysis on the Chinese corpus. Two cases are discussed as follows:

Case 1: Suppose that W_r has no ambiguity, that is, W_r has only a single sense c_r. Furthermore, add extra constraint, suppose that each sense c is derived from different word W, then Equation (4) can be written as:

$$\Phi(c_i' \mid W_r) = \Phi(c_i' \mid c_r) = \frac{\Phi(c_i' c_r)}{\Phi(c_r)} \quad (5)$$

Let $Freq(c)$ denotes the occurrence density of sense c in the Chinese corpus, $Freq(c_i' c_r)$ and $Freq(c_r)$ denote

the occurrence density of correlation sense $c_i'c_r$ and c_r in the corpus. Equation 5 can be approximated:

$$\frac{\Phi(c_i'c_r)}{\Phi(c_r)} = \frac{Freq(c_i'c_r)}{Freq(c_r)}$$

The most reasonable sense of W_i is

$$\begin{aligned} c_{im} &= \arg\max_{c_i'} \Phi(c_i' \mid W_r) = \arg\max_{c_i'} \frac{Freq(c_i'c_r)}{Freq(c_r)} \\ &= \arg\max_{c_i'} Freq(c_i'c_r) \end{aligned}$$

The above equation indicates that the sense of W_i should be chosen the one whose correlation with sense of W_r appears most frequently in the corpus.

Case 2: W_r also has ambiguities. From context analysis, W_i is W_r's relation word, too. They are relation words with each other. Still suppose that each sense *c* is derived from different word *W*. In this case, the probability model (Equation 4) is transferred into dualistic model:

$$(c_{im}, c_{rm}) = \arg\max_{(c_i', c_r')} \Phi(c_i', c_r' \mid W_i, W_r)$$

In which,

$$\Phi(c_i', c_r' \mid W_i, W_r) = \frac{\Phi(c_i'c_r')}{\sum_{c_i, c_r} \Phi(c_i c_r)}$$

From above two equation, lead to:

$$\begin{aligned} (c_{im}, c_{rm}) &= \arg\max_{(c_i', c_r')} \Phi(c_i', c_r' \mid W_i, W_r) \\ &= \arg\max_{(c_i', c_r')} \frac{\Phi(c_i'c_r')}{\sum_{c_i, c_r} \Phi(c_i c_r)} \\ &= \arg\max_{(c_i', c_r')} \Phi(c_i'c_r') \\ &= \arg\max_{(c_i', c_r')} Freq(c_i'c_r') \end{aligned} \quad (6)$$

This illustrates that the senses of W_i and W_r are determined by their most frequent occurrence of correlative sense.

Summarize the above two cases, the following steps are needed to select the proper sense.

1. Determine the relation word of W_i through syntactic analysis;
2. Compute the occurrence density of each correlative sense, and select the sense that makes up the most frequent correlative sense.

2.6. Revision of Disambiguation Model

The above analysis takes the presupposition that each sense *c* is derived from different word *W*. But actually this is not always true. So the model should be revised to fit the real situation. Suppose sense *c* may be derived from the word W', which may take value from the word set:

$$W' \in [W_{p1}, W_{p2}, \ldots, W_{pm}].$$

Rewriting the dualistic model as:

$$\begin{aligned} (c_{im}, c_{rm}) &= \arg\max_{(c_i', c_r')} \Phi(c_i', c_r' \mid W_i, W_r) \\ &= \arg\max_{(c_i', c_r')} \frac{\Phi(c_i', c_r', W_i, W_r)}{\Phi(W_i, W_r)} \\ &= \arg\max_{(c_i', c_r')} \Phi(c_i', c_r', W_i, W_r) \end{aligned}$$

Because the correlative sense (c_i', c_r') may be derived from any correlative word (W_i', W_r'), in which W_i' can be any word that has sense c_i', W_r' can be any word that has sense c_r'. Then $\Phi(c_i', c_r', W_i, W_r)$ can be estimated by the following equation:

$$\Phi(c_i', c_r', W_i, W_r) = \Phi(c_i', c_r') \times \frac{\Phi(W_i, W_r)}{\sum \Phi(W_i', W_r')}$$

If $\Phi(W_i', W_r')$ is estimated by the occurrence density $Freq(W_i', W_r')$ in English training corpus, and $\Phi(W_i, W_r)$ is estimated by $Freq(W_i, W_r)$. Combine the above two equations into the following equation:

$$\begin{aligned} (c_{im}, c_{rm}) &= \arg\max_{(c_i', c_r')} \Phi(c_i', c_r', W_i, W_r) \\ &= \arg\max_{(c_i', c_r')} \frac{Freq(c_i'c_r') \times Freq(W_i, W_r)}{\sum Freq(W_i', W_r')} \end{aligned} \quad (7)$$

From this equation, in general case, steps of sense disambiguation by using bilingual corpus are as follows:

1. Determine the combination word W_r of the polysemous word W_i;
2. Select the sense according to equation 7.

3. Algorithm of Sense Disambiguation

The previous section describes the mathematical model of WSD. This section gives the detail algorithm realization of rule learning and rule application based on the model.

3.1. Model of Rule Learning Rule Template

In order to make the learned rules not only fit the context of the individual words, but also suit the similar context, the rule template contains the morphology, POS and sense information of context words.

Rule template takes the following form:

SELECT Sense_Item IF

a) Match_Word(Context)

b) Match_POS(Context)

c) Match_Sense(Context)

Where "Context" indicates the phrasal collocation, "Word", "POS" and "Sense" denote the word morphology, POS and sense feature respectively. Step *a*, *b* and *c* mean to calculate the similarity in three aspects of morphology, POS and sense.

Weighting Function

From the previous analysis, given relation word pair (W_i, W_r), define the weight of each correlative sense pair (c_i', c_r') as equation 8.

$$R = Freq(c_i'c_r') \times \frac{Freq(W_i, W_r)}{\sum Freq(W_i', W_r')} \quad (8)$$

The sense pair that makes the above expression maximum should be selected.

Algorithm of Rule Learning

Algorithm 1: Algorithm of rule learning

BEGIN

1) Determine the relation word W_r of the polysemous word W_i through context analysis;
2) Generate the correlative sense pair $[(c_i', c_r')]$ from the relation word pair (W_i, W_r);
3) Calculate the weight value R of each correlative sense pair according to equation 8;
4) Select the sense pair (c_{im}, c_{rm}) that make a maximum weight R;
5) Generate a WSD rule and add it into rule list;
6) Apply this rule to the English training corpus;
7) Go to step 1.

END

Give an example to explain the sense selection procedure:

Example: I drive a car to the lake.

The word "drive" has two senses after POS tagging: drive: *v.* 驾驶(control a vehicle), 驱赶(force animal to go);

Determine the relation word W_r ="car";

Construct the correlative sense pair: [(驾驶汽车), (驱赶汽车)];

Calculate the weight *R* of the two correlation pairs and select the element (驾驶汽车) that makes *R* maximum value;

Select the proper sense: drive: *v.* 驾驶(control a vehicle).

3.2. Algorithm of Sense Disambiguation

The learned rules can be directly applied to the sense disambiguation. But in the experiment we found that although some polygamous words' ambiguity types do not match rules of the sense disambiguation rule list, these words can also be disambiguated by the rule list. Algorithm 2 describes the sense disambiguation steps.

Algorithm 2: Algorithm of sense disambiguation

BEGIN

1) Do POS disambiguation on the English training corpus;
2) Extract one polysemous word W_i, and determine its relation word W_r;
3) Find the rule that matches the W_i and its context, if found, complete the sense disambiguation of W_i, and go to step 2; otherwise continue to the next step;
4) Weaken the rule matching condition and find the similar rule again, if found, complete the disambiguation of W_i;
5) Go to step 2.

END

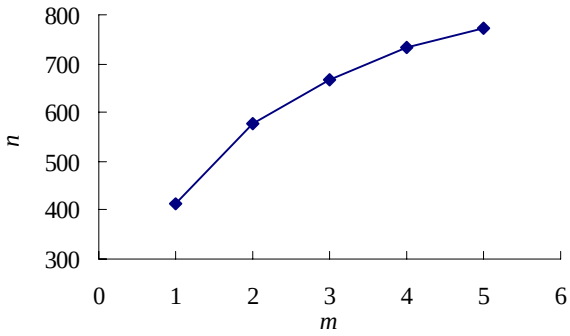
Step 4 weakens the rule matching condition in order to increase the rule's adaptability to fit the similar context, and increase the rule's recall ratio at the same time.

4. Experiments

This section gives the analysis on the WSD method. To examine the effect of the algorithm, a test is directed to compare the WSD result with two other methods, and the result is encouraging.

4.1. Rules Learning

To make the corpus a high coverage, the authors choose the Chinese "Reader's Digest" as Chinese corpus, which has about 10 million Chinese words; and chooses "New Concept English" and some other English text about 60 thousand words as English corpus. In order to appraise the effect of the learned rules, fifty papers are prepared for the open test.



Graph 1: Relationship between lenghth of training text and number of learned rules

Graph 1 gives the relationship between word number of English training text and the number of learned rules, in which *m* represents the number of words of English training text, in unit of ten thousand; *n* represents the number of learned rules. It can be seen from the graph that the number of learned rules tends to level off as the word number of training corpus increases.

Here two examples of the learned rules for the word "fire" are given:

- Select (fire/*v*/开) If Word(W_i)="fire",
  Word(W_r)="gun" OR (POS(W_r)=noun AND SENSE(W_r)='y')
  eg: The criminal fired(开/解雇) the pistol(noun,'y').
- Select (fire/*v*/解雇) If Word(W_i)="fire",
  Word(W_r)="worker" OR (POS(W_r)=noun AND SENSE(W_r)='h')
  eg: The boss fired(开/**解雇**) him(noun,'h').

4.2. Experiment Result

The prepared text is used to test the learned rules. The two parameters that are very important in nature language processing are adopted in appraisement. Recall *r* and Precision *p*. Let *a* represent the number of polysemous words in the test text, *b* represents the number of words whose sense ambiguities are removed by the rules, *c* represents the number of correctly disambiguated words. Then *r* and *p* are given by $r=b/a$, $p=c/b$.

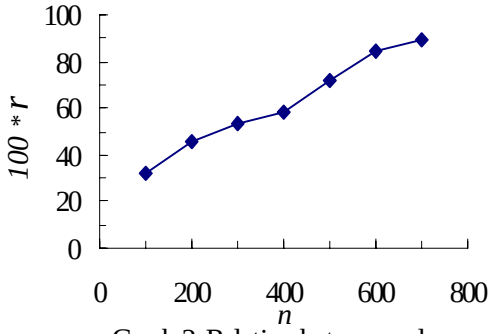
The test is divided into three groups. Each group contains rules that are learned by different methods. The

first group uses the maximum possibility method. This method directly selects the sense of W_i that occurs most frequently in Chinese corpus. It does not taken into account the context of W_i. The second group uses the unrevised model given by equation 5. The third group uses the algorithm 1 after the model has been revised. The result is shown in Table 2.

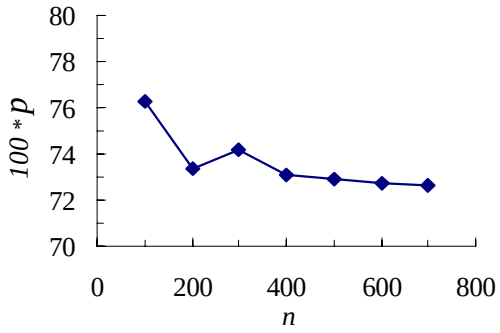
	First Group	Second Group	Third Group
Number of polysemous words	3175	3175	3175
Number of disambiguated words	3175	2981	2975
Number of correctly disambiguated words	1816	2073	2147
Recall (r)	100%	93.8%	93.8%
Precision (p)	57.1%	69.5%	72.2%

Table 2: Test Result

From the table, it can be learned that the third group, which uses the rules learning method of algorithm 1, increases the overall precision by 10.6% compared with the first group, which uses the maximum possibility method. (93.8%*72.2%-100%*57.1%= 10.6%). The third group has the same recall value with the second group, but its precision is a little higher. This is consistent with the mechanism of the two models. They generate the same rule condition while they may bring on different result.



Graph 2: Relation between rule number and recall



Graph 3: Relation between rule number and precision

Graph 2 and graph 3 describe the influence of learned rule number on the recall and precision respectively, in which n indicates the number of learned rules, r represents the recall and p represents the precision.

4.3. Result Analysis

Through the analysis of the test results, the problem in following respects causes the sense disambiguation error.

1) The sense disambiguation error on preposition or conjuncture is fairly high, because these kinds of words have a much more flexible usage than other kinds of words.
2) Weakening the rule condition method is adopted to increase the recall ability, but simultaneously it decreases the precision.
3) Some words combine with other words to form a phrase, which may have no corresponding meaning in Chinese by itself. Such as "take a photo". This kind of words will be correctly disambiguated in phrase process procedure.

5. Conclusion

The paper proposes a method to select the most reasonable sense of a polysemous word in machine translation. It utilizes the English and Chinese bilingual corpus to determine the sense. The key of the method is to determine the word's relation word. Since this method does not require the bilingual corpus to be parallel-aligned, the corpus is easy to acquire. This method has been adopted by our English-Chinese machine translation system and has shown a good result.

6. References

Brill, E. 1995. Unsupervised learning of disambiguation rules for part-of-speech tagging. In *Proc. of 3rd Workshop on Very Large Corpora*, page 1-13. Cambridge, Massachusetts, USA: Association for Computational Linguistic.

Church, K. W. Char-align: A Program for Aligning Parallel Texts at the Character Level. In *Proc. of the 31st Annual Conference of the Association for Computational Linguistics.*

Ide, Nancy. 1998. Introduction to the Special Issue On Word Sense Disambiguation: The State of the Art. In: *Computational Linguistics*, volume 24, number 1, pages 1-38.

Ker, S. J. 1997. A Class-based Approach to Word Alignment. *Computational Linguistics*, Volume 23, Number 2,

Liu, X. H. 1997. Sense Selecting Based on Corpus. *Information Journal*, volume 16, number 3, page 189-194.

Li, J. Z. 1999. Sense Tagging Method Based on Unsupervised Transformation. *Journal of Tsinghua University*, volume 39, number 7, page 116-120.

Yu, S. W. 1993. "Computer Linguistics". China, Peking Univ. Press.