International Journal of Knowledge and Language Processing
Volume **2**, Number 2, April **2011**

www.ijklp.org
AKLP International ©2011 ISSN 2191-2734
pp. **35-43**

Bilingual Words Sense Disambiguation in English-Chinese Parallel Corpus

Feng Min-xuan[1] Qu Wei-guang[2,3]

[1]School of Chinese Language and Literature
Nanjing Normal University, Nanjing 210046, China
fennel_2006@163.com

[2]School of Computer Science and Technology
Nanjing Normal University, Nanjing 210046, China
wgqu_nj@163.com

[3]The Research Center of Information Security and Confidentiality Technology of Jiangsu Province
Nanjing, 210097, China

Received April 2011; revised April 2011

ABSTRACT. *This paper reports experiments of idiosyncratic-rule based WSD in parallel corpus, aiming at solving those highly frequent polysemes which are poorly handled in statistical methods. We have compiled some idiosyncratic rules which are not confined by context length or number of rule templates and are fit for bilingual parallel processing. The parallel processing algorithm designed for the 5 polysemes (地方、所有、等、since、state) are tested and validated in large scale parallel corpora. The tagging precision rates in observation corpus all reaches 100%. In other non-observation corpora, the F-score varies between 90% and 100% which is much better than the state of art of monolingual WSD performance.*

Keywords: parallel corpus, polysemy, word sense disambiguation, automatic recognition, Chinese information processing

1. **Introduction.** The most frequently used method in Machine Translation is to the corpus-based selection of translation. Based on bilingual corpus, Zhang Jing (2003) uses unsupervised learning methods to select translation and achieves better results than methods based on maximum probability[1]. Liu Dongming et al (2005) makes use of HowNet and sentence-level Chinese-English bilingual parallel corpus and convert the problem of word sense disambiguation into the problem of semantic similarity calculation of sense combination between corresponding sentences in two languages. The research reports a highest precision of 90.8% for Chinese and 92.4% for English [2]. To further enhance the sense tagging accuracy of both Chinese and English words in parallel corpora, we propose to formulate word sense disambiguation methods based on idiosyncratic rules for those words which are highly frequent but poorly dealt with in current statistical methods.

2. Polysemy in Parallel Corpus. According to whether the word belongs to different part of speech or not, a polysomic word can be categorized into two classes: Solo-POS word and Multi-POS word. In the corpus, the information of the words' part of speech is encoded in POS tags. We calculate the distribution of 4724 (25796 occurrence) Chinese words and 5144 (24,407 occurrence) English word in the PCCE1000[1] as is shown in Table 1.

TABLE 1. Statistics of Polysemy in PCCE1000

Category	Type Num	Frequency Percentage(%)	Token Num	Frequency Percentage(%)
Chinese Solo-POS	169	3.58	1073	4.16
English Solo-POS	1992	38.72	8229	33.72
Chinese Multi-POS	138	2.92	4874	18.89
English Multi-POS	201	3.91	2765	11.33

The so-called Solo-POS polysemes refer to those words in the corpus which occur with just one type of part of speech tag. These words include the class of words with only one type of part of speech, such as "材料(material), deliver, importance " and the class of words which have more than one type of speech but occurs only with one type of part of speech due to limits of corpus size, such as "打(hit), 高度(height), cover " and so on.

The so-called Multi-POS polysemes[2] refer to those words in the corpus which occur with more than two types of part of speech, among which are mainly Multi-POS-spanning polysemes, for example "把（p、q）"3, "关系（n、v）", "根本（a、d）", "face（NN、VB）", "out（IN、RB）", "work（NN、VB、VBP）" etc.

From the statistics in Table 1, we find that there are more polysemes in English than in Chinese. This conclusion is similar to previous studies [3]. But the majorities are solo-POS polysemes, indicating that although many English polysemes may have multiple parts of speech, but in actual corpus, they often occur with one part of speech due to limitations on corpus size and content. Therefore, the complexity of elimination of Chinese and English polysemy is still similar.

3. Selection of Target Words for Idiosyncratic WSD. To make the work of idiosyncratic WSD in parallel corpus worthwhile, the target words much be carefully chosen by taking into considerations the frequency, word class, difficulty in processing and other aspects. As a result, three Chinese polysemes , namely "地方", "所有", and "等" and two English polysemes, namely "since" and "state" are chosen to verify the feasibility and effectiveness of the parallel processing algorithms [4]. The following gives an analysis on the characteristics of these polysemes.

Firstly, "地方" is a solo-POS polyseme, which is defined as below:

1. PCCE1000 is a bilingual corpus compiled by Language Technology Lab, Nanjing Normal University. The corpus contains 1000 beads, of 82 articles of news.

2. In English, the base form and non-third singular form of some verbs, or the past form and past participant form are identical in word form. resulting in the taggs of VB and VP, or VBD and VBN. "know", "think" and "pleaded" are such examples. These words are also regarded as Multi-POS polysemes. The tags inside the brackets are POS taggs in PCCE1000.

《ABC 汉英大词典》（2003）[5]

di4 fang1

1:local；regional　　2.place；site；locality

di4 fang5

1.place;space;room　　2.territory

《现代汉语词典》（2002）[6]

[地方]di4fang1①各级行政区划的统称 （跟‘中央’相对）；中央工业和～工业同时并举。②本地；当地：他在农村的时候，常给～上的群众治病。

[地方]di4fang5 ①（～儿）某一区域；空间的一部分；部位：你是什么～的人？｜你听，飞机在什么～飞？｜会场里人都坐满了，没有～了｜我这个～有点疼。②部分：这话有对的～，也有不对的～。

It can thus be seen that the two “地方” in the dictionary are of the same word-form and part of speech. In WSD, four senses must be considered:（1）name for different levels of administrative regions;（2）local place;（3）a certain region or part（4）part. Accordingly, we have adopted the policy to consider the two “地方”as one word with different senses in WSD. In parallel corpus, they are frequently found. In “English Online”, a total of 741 beads are found with this word. [7]

Secondly, the word “所有” is also a content polyseme, but it belongs to Multi-POSes polyseme. The definition of the word is given below:

《ABC 汉英大词典》（2003）

1.v.　own;possess

2.adj.　all;every;all there are

3.n.　possession

《现代汉语词典》（2002）

[所有]suo3you3①领有：～权|～制。②领有的东西：尽其～。③一切；全部：把～的力量都贡献给祖国。

According to analysis on corpora and dictionary, the word “所有” has 3 senses when used as an independent word: (1) to possess ; (2) possession; (3) all. These three senses are used as verb, noun and distinguisher respectively. Thus this word falls into the category of one-sense-one-part-of-speech type, and the problem of word sense disambiguation is

equivalent to part of speech disambiguation. We have conducted statistics on part of speech tagging[8] on PCCE1000, among which the 16 occurrences of the word “所有” (2 occurrences are verbs and 14 occurrences are distinguishers), are all tagged with b (tag for distinguisher). In addition, the word “所有” can be used together with “人” to form a combinatory ambiguous string “所有人”. These POSes are difficult to handle when the WSD is handled within one language. In addition, in the “English Online “, a total of 1243 sentences are found, showing that the frequency is very high.

Finally, “等” is a POS-spanning polyseme, which is also of high frequency. In the “English Online”, a total of 1077 sentences are found. Analysis on corpora and dictionary definition shows that “等” has four senses which used independent as a word: (1) level; (2) is equal to; (3) wait; (4) and so on. Therefore, it can be used as a noun, a verb (with two senses), or an auxiliary. Thus this word is both a content word and a functional word and falls into one-part-of-speech-multiple-sense type. In other words, if we can solve the WSD problem of the word, the part of speech ambiguity can also be solved. According to statistics, PCCE1000 has a total of 10 occurrences of “等” (two occurrences of noun, 5 occurrences of verb, and 3 occurrences of auxiliary), of which two occurrences of noun and four occurrences of verb are tagged as u (auxiliary). The part of speech disambiguation is thus poor and cannot be used for reference. In addition, the word “等” combines easily with other words to form ambiguous strings. For example, “上等”, “下等” are ambiguous, “等于是” is also wrongly segmented, “中等”, “低等” are also wrongly segmented, “高度” is regarded as one word. To sum up, “等”“ is not handled successfully at the level of word segmentation and POS tagging, which makes WSD even more difficulty.

《ABC 汉英大词典》（2003）

v.（动词）wait

B.f.（黏着词素）①class,grade②be equal③division

SUF.（后缀）and so on

《现代汉语词典》（2002）

等[1] deng3 ①等级：同～｜优～｜共分三～。②种；类：这～事｜此～人。③程度或数量上相同：相～｜～于｜大小不～。④同‘戥’（deng3）。

等[2] deng3 ①等候；等待：～车｜请稍～一会儿｜～他来了一块儿去。②等到：～我写完这封信再走也不晚。

等[3] deng3 助词。①（书）用在人称代词或指人的名词后面，表示复数：我～｜彼～。②表示列举未尽（可以叠用）：北京、天津～地｜纸张文具～～。③列举后煞尾：长江、黄河、黑龙江、珠江～四大河流。

The English word “since” is also both a content word and a functional word, and a polyseme spanning different types of POSes. In the “English Online”, a total of 442 sentences are obtained.

> *《朗文当代高级英语辞典》（2004）*[9]
>
> *conj.从...以来，从...以后；从...以后（的一段时间里），自从...以来一直...；*
>
> *因为，既然*
>
> *prep.自从...以来，自从...之后；自从...以后（的整段时间里），从...以后一直...*
>
> *adv.从那时以来， 后来；从那时至今（的整段时间里），从那时起一直*
>
> *《牛津高阶英汉双解词典》（2002）*[10]
>
> *prep.从（过去某时间）以来、以后或到现在*
>
> *conj.从（过去某事）以来、以后或到现在；因为；既然；由于*

Form the definition given above, it can be seen that “since” has three types of POSes, namely preposition, conjunction and adverb. When used as a preposition or a conjunction, it means “from that time onward”. Some additional senses are also conveyed by the word when it is used as conjunction. Thus, this word belongs to the “multi-senses-with-multi-part-of-speeches” type which is very complicated. In PCCE1000, 27 occurrences of “since” are all tagged with IN (part of speech tag for preposition or conjunction). Coupled with the fact that “since” is mainly used as functional word, it is difficult in monolingual WSD.

The English word “state” is a functional word with more than one types of POS, it is also of the “multi-senses-with-multi-part-of-speeches” type. The definition of the word is given below:

《朗文当代高级英语辞典》（2004）

n.状况，状态，情况，情形；政府；国家；国家的一部分，州，邦；盛礼，隆重的仪式

vt.陈述，说明；（文件、报纸、票据等）写明；

《牛津高阶英汉双解词典》（2002）

n.状态；状况；情况；情形；国；国家；领土；（联邦或共和国的）州，邦；政府；国家；国家级的礼仪；盛观

adj.政府的；国家的；关于国家的；礼仪的；礼仪上的；礼节性的

v.陈述或说明（某事）；预先安排、定下或宣布（某事）；规定；确定

It can be seen that "state" can be used as a noun, an adjective or a verb. When used as a noun or an adjective, it has the sense of "country, state or government". When used as a verb, it also has the sense of "country, state, government". Thus it is a content polyseme which spans different types of POSes. The word can also be found in "English Online" , with a total of 202 sentences.

4. The Algorithm of Parallel WSD. The basic idea of parallel polysemy disambiguation is to obtain knowledge template based on dictionary analysis and corpus examination, together with introspective speculation. Then the rules are used to tag the senses of the polysemes in the sentence-level parallel corpus.

First of all, the WSD algorithms for polysemy "地方" and "所有" are similar. For "地方", the task is to sense-tag the word in the parallel corpus with the following senses: (1) levels of administrative division in general; (2) local; (3) a region or part of a region; (4) part. For "所有", the task is to sense-tag the word in the corpus with following senses: (1) possess; (2) possession; (3) all. The algorithm is as follows:

(a) Find sentence beads in the corpus with the word "地方" or "所有", and record the number of occurrence of the word in each sentence. If the word is found, move on to step b, otherwise move to finish;

(b) In the sentence bead found with occurrence of the word("地方" or "所有"), look for the matching English translation in the English sentence. For example, the word "地方" has several different corresponding English translation. These different versions of translation can be used to determine the sense of the word: if the translation is "district", the sense is determined to be "levels of administrative division in general"; if the translation is "place", the sense is determined to be "a region or part of a region".

(c) If the sentence bead is found to have more than one occurrence of "地方", mark each occurrence of the word according to "ground level" from the highest to the lowest. If no matching English translation is found for the word, assign to the word the most frequent sense (Sense Three), namely "a region or part of a region". The same procedure can be applied to the word "所有". The most frequent sense for this word is also the third sense:

"all".

For the word "等", the most frequent sense is "wait", one of the four senses (please refer to Section II). The WSD algorithm for the word is similar, except that in the second step, a further singular language WSD is applied. For example, the word "等" in the phrase "等一等" should be tagged with the sense "wait".

The WSD for the English Polyseme "since" needs to distinguish between 5 senses (see section II). The WSD for "state" needs to distinguished between 8 senses (also see section II). The three-step process used for the Chinese words can also be applied here with some minor changes: in first step, when recording the occurrence of the word "since", the occurrence of "after", "left" and "then" should also recorded; when recording the occurrence of the word "state", the occurrence of it synonyms such as "national", "country", "government", "union", "state-owned" and "provincial" should also be recorded, in order to avoid interference; in the third step, the most frequent sense for "since" is the first sense, the most frequent sense for "state" is the 6th sense.

5. Experiment Results. We use 3 parallel corpora for the parallel polysemy disambiguation test:

(1)Type I: PCCE1000, which is used as observation corpus4, with 1000 beads of legal news and information and intelligence.

(2)Type II: English-Chinese parallel corpus, obtained from "Chinese line". From which about 50 sentence beads are taken for each word in question. The corpus is diverse in subjects. This is used as the non-observation corpus and used for test.

(3)Type III: EWS corpus, with 4198 beads, subjects focused on news. This corpus is also used for test.

Our algorithm conducts WSD on the 5 polysemes mentioned. The results are given in Table 2. Meanwhile, the senses of the polysemes are not evenly distributed in the corpora. Table 3 and Table 4 give out the frequency and sense tagging for the words in question.

TABLE 2. Experiment Results for the Three Types of Corpora

Words	Corpus Type I		Corpus Type II		Corpus Type III		Sum	
	Num	F（%）	Num	F（%）	Num	F（%）	Sum	F（%）
地方	50	100	19	94.74	48	97.92	117	98.29
所有	48	100	17	100	119	100	184	100
等	55	100	10	90	111	95.50	176	96.59
since	51	100	27	100	55	96.36	133	98.50
state	50	100	24	100	97	96.91	171	98.25

[4] Observation corpus refers to the corpus used for analysis and development of disambiguation rules. Besides dictionaries, observation corpus is the most important tool for WSD rules. A large number of introspective disambiguation rules are derived from the analysis of the observation corpus.

TABLE 3. Sense Distribution For Chinese Polysemes

Sense Index	地　方			所　有			等		
	Num	Percentage（%）	F（%）	Num	Percentage（%）	F（%）	Num	Percentage（%）	F（%）
（1）	28	23.93	96.43	2	1.09	100	8	4.55	100
（2）	11	9.40	90.91	0	0	100	12	6.82	83.33
（3）	68	58.12	100	182	98.91	100	42	23.86	100
（4）	10	8.55	100				114	64.77	96.49

TABLE 4. SENSE DISTRIBution for English Polysemes

Sense Index	since			state		
	Num	Percentage（%）	F（%）	Num	Percentage（%）	F（%）
（1）	70	52.63	100	75	43.86	97.33
（2）	30	22.56	96.67	7	4.09	100
（3）	20	15.04	95.00	23	13.45	100
（4）	10	7.52	100	2	1.17	100
（5）	3	2.26	100	13	7.60	92.31
（6）				46	26.90	100
（7）				0	0	100
（8）				5	2.92	100

6. **Conclusions.** The Rule-based methods need to consider the following questions: (1) the coverage of the rules, namely how widely can a rule be used; (2) the correct ratio, namely the ratio between the times of correct WSD and the times of application of the rule; (3) the order of application, which needs to be addressed in order to avoid collision between rules. In rule extractions, we have taken into account all these problems. In Particular, to address the third problem, we have adopted a "sort-based-on-ground" approach to achieve an appropriate order of rules.

Acknowledgment. This work is supported by Jiangsu Social Science Fund(10YYB007）, Chinese National Science Fund(60773173, 61073119), Chinese National Social Science Fund(07BYY050,10CYY021）, Jiangsu Science Fund（BK2010547）, the Fund of Social Science of the Ministry Department in Jiangsu Province under Grant 06SJB71007 and the third Principle Research Fund of "211 Project" in Nanjing Normal University "Language Technology Innovation and Construction of Working Platform".

REFERENCES

[1] Zhang Jing. Unsupervised Translation Selection Based on Non-parallel Bilingual Corpora. *Harbin industrial university Ph.D. Thesis*, pp.81, 2003.

[2] Liu Dongming,Yang Erhong,Fang Ying. Word Sense Tagging in Chinese- English Parallel Corpora. *Journal of Chinese Information Processing*, vol.19, no.6, pp.50-56, 2005.

[3] Wu Fengjuan. Contrastive Study on Ambiguty in Chinese and English. *Huazhong University of Master Degree Theses*, pp.34, 2004.

[4] Feng Minxuan. Parallel Processing on Parallel Corpus of Chinese-English. *Nanjing normal university PhD thesis*, pp.9, 2006.

[5] John Defrancis. *ABC Chinese-English Dictionary*. Shanghai: Chinese Big Dictionary Press, 2003.

[6] Institute of the Chinese Academy of Templates Language Dictionary. *The Contemporary Chinese Dictionary (greenblatt)*. Beijing: The Commercial Press, 2002.

[7] Lu Wei . *Chinese-English Parallel Corpus* . http://www.luweixmu.com/ec-corpus/query.asp,2010.

[8] Institute of Computing Technology Chinese Academy of Sciences. *ICTCLAS*. http://www.duanxinhui.com/soft/html/3861.html,2010.

[9] British Pearson education publishing Co., LTD. *Longman Dictionary of Contemporary English*. Beijing: Foreign Language Teaching and Research Press, 2004.

[10] Oxford University.*Oxford Advanced Learner's English-Chinese Dictionary(Revised Extended 4th Edition)*.Beijing: The Commercial Press, 2002.