Sense Prediction Study: Two corpus-driven linguistic Approaches

Jia-Fei Hong[1] Sue-Jin Ker[2] Kathleen Ahrens[1,5] Chu-Ren Huang[3,4]

[1] Graduate Institute of Linguistics, National Taiwan University, Taiwan

[2] Department of Computer Science and Information Management, Soochow University, Taiwan

[3] Institute of Linguistics, Academia Sinica, Taiwan

[4] Faculty of Humanities, The Hong Kong Polytechnic University, Hong Kong

[5] Language Centre, Hong Kong Baptist University, Hong Kong

E-mail: jiafei@gate.sinica.edu.tw; ksj@cis.scu.edu.tw

kathleenahrens@yahoo.com; churenhuang@gmail.com

Abstract: In this study, we propose to use two corpus-driven linguistic approaches for a sense prediction study. We will concentrate on the character similarity clustering approach and concept similarity clustering approach to predict the senses of non-assigned words by using corpora and tools, such as Chinese Gigaword Corpus, and HowNet. In this study, we would then like to evaluate their predictions via the sense divisions of Chinese Wordnet (CWN) and *Xiandai Hanyu Cidian* (*Xian Han*). Using these corpora, we will determine their clusters of our four target words ---- *chi1* "eat", *wan2* "play", *huan4* "change" and *shao1* "burn" in order to predict their all possible senses and evaluate them. This requirement will demonstrate the visibility of the corpus-based approaches.

Keywords: Lexical ambiguity; Sense prediction; Corpus-based approach; Character similarity clustering approach; Concept similarity clustering approach; Evaluation

1 Introduction

Our goal in this study of sense prediction is to generate solutions for lexical ambiguity in general. In particular, we will look at words without lexically-assigned senses and try to predict the range of senses each word form may have. Since lexical information of senses of these words is not available, we propose to use corpus-driven distribution as the main information is prediction. We will determine the collocation clusters of our target word through characters, semantic features, and concepts by using corpora and tools, such as Chinese Gigaword Corpus, Chinese Wordnet and *Xiandai Hanyu*. Our study showed the feasibility of sense-prediction without lexically assigned senses.

Lexical ambiguity is a linguistic term for a word's capacity to carry more than two senses at the same time, for example, *bank*. In some modern linguistic and literary theories, the term has been extended to larger units, including entire literary works. Several previous studies concerning lexical ambiguity seem very popular. In the case of the previously related lexical ambiguity, several studies have concentrated on the corpus-based and computational perspective included: Peng et al. (2007), Xue et al. (2006), Chen et al. (2005), Moldovan and Novischi (2004) …and so

on and they took several different approaches: used the corpus-based approach, an adaptive system, divided the sense of lexically ambiguous word and found the possibility of the senses of a word.

We know of several researchers who use only manual analysis to find out the argumentative roles and predict their semantic features to determine their senses. Therefore, they can't deal with more quantities of lexically ambiguous words at the same time. We consulted Fujii and Croft's study (1993) to collect related collocations to categorize different clusters by using the character similarity clustering approach for achieving automatic sense prediction. In addition, we also consulted Liu and Li (2002), Li et al. (2005) and Dai et al. (2008) to take different dimensions to calculate and obtain the similarities by using the concept similarity clustering approach.

Overall, regarding these previous corpus and computational studies, these scholars proposed corpus-based, algorithm, automatic programming system, and collocation approaches to analyze sense prediction studies. Unfortunately, they only employed one corpus in their studies and got less information of lexical ambiguity; they also did not combine these approaches.

2 Research Question

When we define lexically ambiguous senses, we need to notice that (1) senses are represented as sets of necessary and sufficient conditions that fully capture the conceptual content conveyed by words; (2) there are as many particular senses for a word as there are differences in these conditions; and (3) senses can be represented independently of the context in which they occur.

Regarding lexical ambiguity, there are two hypotheses in this study. First, lexical ambiguity is the property of some words to have multiple meanings or senses (Moldovan and Novischi, 2004). We would like to follow Fujii and Croft (1993) to observe the character similarity and refer to Li et al. (2003) and Dai et al. (2008) to explore the concept similarity by using HowNet. Second, collocation was a combination of words that has a certain tendency to be used together, it was used widely to attack the WSD task and word classes were often used to alleviate the data sparseness in NLP (Peng et al., 2007) Therefore, using the argument role, labeling information can help us to extract some types of semantic features. For this reason, there are two research questions in this study: (1) How do we predict the word senses of a lexically ambiguous word to present different interpretations in different contexts or domains? And (2) How do we use a corpus as the database to support a word sense prediction study?

3 Analysis

3.1 Methodology

In this sense prediction study, we would like to explore all possible senses of my four target

words --- *chi1* “eat”, *wan2* “play”, *huan4* “change” and *shao1* “burn”, therefore, we need to collect large data to analyze and examine them in order to achieve the objectivity, equitable and rational. Chinese Gigaword Corpus is a good candidate. In addition, it’s because we will map and assign all related collocation words of four target words in the concept similarity clustering analysis to represent their semantic features and concepts, we choose HowNet as the knowledge bases. When we do the sense prediction study, of course, it’s necessary to evaluate their evaluations. We then select CWN and *Xian Han* as criteria to evaluate for these four target words.

In order to collect large data to explore our sense prediction study, we focus on Taiwan’s Central News Agency Gigaword Corpus. The Chinese Gigaword Corpus contains about 1.4 billion Chinese characters, including about 800 million characters from Taiwan’s Central News Agency (from 1991 to 2004), nearly 500 million characters from China’s Xinhua News Agency, and approximately 30 million characters from Singapore Zaobao.

We would like to present their semantic features and concepts of the related collocation words for these four target words in the concept similarity clustering approach. HowNet is the knowledge bases which can show their internal semantic components, features and combination of sememes and pointers for all words in detail. HowNet includes an abundance of both semantic and world knowledge and thus is an important resource for NLP and knowledge mining.

For the evaluations of the four target words, we will use CWN and *Xian Han*. The architecture of CWN follows the standard established by Princeton’s WordNet, which has two unique design features. Huang (2003) proposed the criteria and operational guidelines for the process of dividing lexical senses. In addition, we then take *Xian Han* (the fifth version, 2005) to evaluate them in the same time. In terms of *Xian Han*, it’s the Modern Chinese Dictionary which is the first Chinese dictionary, published by The Commercial Press.

Since the target words are all transitive verbs, their object positions must be nouns; in other words, we can regard these nouns as their important related collocation words. Therefore, we can employ there nouns to predict all possible senses for the four target words. In Chinese, the main object (noun) usually appears after the transitive verb but sometimes the main object (noun) appears before the transitive verb. Following the rules of structural construction in Chinese and in order to gain these related collocations, we use five different ways to collect them: (1) the noun after the target word; (2) the head noun of the first noun phrase after the target word; (3) the head noun of the last noun phrase before the first punctuation mark of the target word; (4) the noun before the first punctuation mark of the target word; and (5) the noun nearest the punctuation mark before the target word. They are shown following, in Table 1.

Table 1: The related collocations

Category	Connected Sentence	Related Collocation
1	民众除了多食用蔬菜，多***吃***鱼也有益健康。	鱼{Na}
2	减肥鸡保证有***吃***五味盐酥鸡的滋味与口感。	盐酥鸡{Na}
3	巴拿马人不***吃***猪内脏与猪脚筋。	猪脚{Na}
4	从***玩***彩色木棒或积木块的游戏中，能轻易学到像是长、高、形状、表面、尺寸。	游戏{Na}
5	蔬菜尽量以凉拌或生***吃***，不加油更佳。	蔬菜{Na}

Following these five selection criteria, there are 29,421 sentences for the collocations of *chi1* “eat”; 8,833 sentences for the collocations of *wan2* “play”; 19,394 sentences for the collocations of *huan4* “change”; and 4,668 sentences for the collocations of *shao1* “burn”. From these sentences, there are 3,961 collocations for *chi1* “eat”; 2,086 collocations for *wan2* “play”; 3,003 collocations for *huan4* “change”; and 1,565 collocations for *shao1* “burn”. This empirical data can then be used to process the character similarity clustering analysis and the concept similarity clustering analysis; moreover, it can predict all of their possible senses.

3.2 Character similarity clustering analysis

Following Fujii and Croft’s study (1993), we will use character similarity to cluster related collocations in order to predict possible senses of the four target words, although by a different method. Similar features are often synonymous compounds that share a common morpheme. For instance, [饭 (*fan4* “rice”), 米饭 (*mi3 fan4* “rice”)] and [案 (*an4* “case”), 案件 (*an4 jian4* “case”)], respectively, share a common morpheme [饭 (*fan4* “rice”)] and [案 (*an4* “case”)]. Fujii and Croft (1993) also pointed out a similar thesaurus effect of Chinese characters in Japanese Information Retrieval. In the cluster step, there are two sub-steps here: (1) character similarity comparison between words; and (2) group similarity comparison between words. Two formulas for these sub-steps are presented as the following:

$$dice(x, y) = \frac{2 | x \cap y |}{| x | + | y |} \tag{1}$$

By using this formula, we will obtain some collocations and regard 药 (*yao4* “medicine”), 减肥药 (*jian3 fei2 yao4* “reducing weight medicine”), and 中药 (*zhong1 yao4* “traditional Chinese medicine”) as the same cluster.

$$sim(x, Y) = \frac{\sum_{y \in Y} dice(x, y)}{| Y |} \tag{2}$$

In Formula 2, 敗績 (*bai4 ji1* “defeat”) and 敗仗 (*bai4 zhang4* “defeat”) can be placed into the same cluster.

After finishing the two sub-steps of the character similarity clustering analysis, we will use another automatic programming strategy to achieve more precise sense clusters by averaging the similarity of two different clusters, as shown in Formula 3 below.

$$sim\ (clu_1, clu_2) = \frac{\sum_{s \in clu_1} \sum_{t \in clu_2} (dice\ (s,t))}{|clu_1| \times |clu_2|} \quad (3)$$

In Formula 3, not only are two similar words clustered into one particular cluster, but also different clusters are combined into clusters with the highest similarity.

In general, observations show that high-frequency words are usually highly ambiguous and have more senses; on the contrary, low-frequency words usually have less senses or only a single sense. Therefore, we attempt to examine these peripheral words individually for these four target words and their frequencies are similar in Taiwan’s Central News Agency of Gigaword Corpus and they have analyzed in CWN already. For example, we can find that the peripheral words of *chi1* “eat” are 造 (*zao4* “produce”, 11 senses in CWN), 按 (*an4* “press”, 9 senses in CWN), 认 (*ren4* “recognize”, 9 senses in CWN), 创 (*chuang4* “invent”, 10 senses in CWN)…and so on. The peripheral words of w*an2* “play”, *huan4* “change” and *shao1* “burn” are all found such as *chi1* “eat”. Therefore, we presume there are 10 senses for *chi1* “eat”, 9 senses for *wan2* “play”, 7 senses for *huan4* “change” and 6 senses for *shao1* “burn”. But before reducing these collocations to these clusters, frequencies that are less or equal two (≦ two) will be cut.

We will focus on more types of clusters to examine their accuracy in 100 clusters, 200 clusters, and 300 clusters for *chi1* “eat”; 90 clusters, 180 clusters, and 270 clusters for *wan2* “play”; 70 clusters, 140 clusters, and 210 clusters for *huan4* “change”; and 60 clusters, 120 clusters, and 180 clusters for *shao1* “burn”. However, in order to achieve an integral, 30 senses of *chi1* “eat”, 10 senses of *wan2* “play”, 6 senses of *huan4* “change”, and 15 senses of *shao1* “burn” will be regarded as the standard default targets. In order to select particular clusters to examine, the testing cluster sizes will be 1, 1.5, and 2 times that of the senses. We not only was able to calculate the accuracy of the sentences and collocation types of the four target words, but also we was able to observe the accuracy of the average distributions, as shown below in Table 2 and Table 3.

Table 2: The precision average distribution of four target words by sentence

	***10**	***20**	***30**
***1**	61.04%	**77.38%**	87.80%
***1.5**	61.73%	**78.05%**	87.20%
***2**	61.87%	**78.45%**	86.86%

Table 3: The precision average distribution of four target words by type

	***10**	***20**	***30**
***1**	51.66%	**59.14%**	70.05%
***1.5**	51.76%	**60.24%**	69.22%
***2**	52.09%	**61.08%**	68.29%

Concentrating on the 20-times prediction clusters, when we set up 20-times predicting clusters as our default targets for the four target words, they indeed followed the reasonable distributions and presented the best results.

3. 3 Concept similarity clustering analysis

Regarding our cluster determination by the character similarity clustering analysis, we concentrate on the same morpheme of all collocations in each cluster. However, if we focus only on the morpheme, perhaps many non-related collocations will be assigned to the same cluster, or perhaps many related collocations will be assigned to different clusters. For example, 山药 (shan1 yao4 “Chinese yam”) and 药 (yao4 “medicine”) are in the same cluster. 汉堡肉 (Han4 bao3 rou4 “hamburger meat”) is categorized into 汉堡 (han4 bao3 “hamburger”) cluster rather than 肉 (rou4 “meat”) cluster.

We would like to attempt to assign all words to lexical concepts via HowNet and we then can calculate their concepts similarities in order to cluster these words. Because HowNet can present more definite semantic elements and semantic features of all words, we will utilize them to examine and ensure feature and concept determination. Owing to more words map to the same concept, they usually are regarded as synonymous words in some kind of degree, for instance, the concepts of *xi1 gua1* “watermelon”, *shi4 zi5* “persimmon”, *ping2 guo3* “apple” and *pu2 tao2* “grapes” are fruits and they are regarded as synonym and clustered in the same cluster.

It’s important view that two main strategies are in concept similarity clustering analysis as 1) similarity between sememes and 2) similarity between concepts through HowNet. HowNet organizes all the sememes into several trees, and each sememe is considered a node of a tree. In this way, we can calculate the distance between any two sememes (Dai et al., 2008). We are also

able to define the distance between the sememes as the length of path between them, as shown in Formula 4.

$$\text{sim\_seme}(s_1, s_2) = \frac{\min(d(s_1), d(s_2))}{\text{dis}(s_1, s_2) + \min(d(s_1), d(s_2))} \quad (4)$$

Following Liu and Li (2002), Li et al. (2005) and Dai et al. (2008), to find the similarity between the two concepts; however, we use three different dimensions to calculate them, sum these three amounts by their weights, and then obtain their similarity. Our schema is shown and expressed in Formula 5.

$$sim\_def(m,n) = \alpha \times sim\_seme(pm, pn) + \beta \times \frac{\sum_i \max_j (sim\_seme(m_i, n_j)}{|m|} + \gamma \times \frac{|m \cap n|}{|m| + |n|} \quad (5)$$

In Formula 5, we can gain the final average similarity in order to determine the sense clusters. In the case of the precisions of these four target words in the concept similarity clustering analysis, it's necessary that we need to examine them manually. We also randomly select some clusters as our testing data. We presume there are 10 senses for *chi1* "eat", 9 senses for *wan2* "play", 7 senses for *huan4* "change" and 6 senses for *shao1* "burn", therefore, we will randomly select 10 clusters for *chi1* "eat", 9 clusters for *wan2* "play", 7 clusters or *huan4* "change" and 6 clusters for *shao1* "burn". After examining these clusters, we can obtain their precision by their sentences. We find out their accuracy rate are all over 84% and the average accuracy rate is 85.90%.

Table 4: Average precision for four target words

Target Word	Accuracy Rate
***Chi1* "eat"**	85.59%
***Wan2* "play"**	87.21%
***Huan4* "change"**	85.98%
***Shao1* "burn"**	84.81%
Average	**85.90%**

When evaluating the sense predictions for the four target words in the character similarity clustering analysis, the data size determined was 20 times the number of sense predictions, and this same data size will be used in the concept similarity clustering analysis. We are able to obtain higher accuracy rates and better performances using the concept similarity clustering analysis of the corpus-based and computational approach in the sense prediction study.

3. 4 Evaluation

After discussing the character similarity clustering approach and the concept similarity clustering approach for the four target words in this study, we will evaluate the performances of the four target words via CWN and *Xian Han*. In CWN and *Xian Han*, the four target words have been analyzed and have been assigned appropriate senses.

Following the principle of calculating the accuracy rates in the character similarity clustering approach, we selected the 2-times number of clusters as our testing data. Based on the character similarity clustering approach, the distributions of the sense prediction evaluations for the four target words in CWN and *Xian Han* are shown in Table 5 below.

Table 5: Evaluations in CWN and *Xian Han* based on the character similarity clustering

Target Word	**CWN**			***Xian Han***		
	Sense	**Tagging**	**Recall**	**Sense**	**Tagging**	**Recall**
***Chi1* "eat"**	28	22	78.57%	8	7	87.5%
***Wan2* "play"**	9	8	88.89%	3	3	100%
***Huan4* "change"**	5	5	100.00%	3	3	100.00%
***Shao1* "burn"**	14	8	57.14%	8	4	50%
Average			**81.15%**			**84.38%**

From Table 5, the evaluations show that some senses cannot be tagged in CWN and *Xian Han* based on using the character similarity clustering approach. *Chi1* "eat", for example, is evaluated in CWN based on the character similarity clustering analysis. We find when we tag sense to these selected character similarity clusters based on sense division in CWN, only 22 senses can be tagged in manual. For example, we can observe CWN sense "使食物经过口中吞入体内 (to take food through the mouth and swallow into the body)" but we can not observe CWN sense "比喻取得对方棋子或牌 (to capture other chesses or playing cards)". At the same time, we also calculate the recall rate. Manually checking the evaluation in *wan2* "play", we observe only one CWN sense can not be tagged which is "没有特定目的用手拨弄后述对象 (toy something by hand purposelessly)". For evaluation in *huan4* "change", we can observe all five CWN senses in manual while for evaluation in *shao1* "burn", we only observe 8 senses in CWN. In the same condition, based on the character similarity clustering analysis, we can not find *Xian Han* sense "被 (多见于早期白话) passive" for *chi1* "eat" in manual and at the same time we either can not find four *Xian Han* senses for *shao1* "burn" such as 发烧 (fever),比正常体温高的体温 (the temperature is higher than normal)... and so on. However, we can observe complete *Xian Han* senses in the evaluations of *wan2* "play" and *huan4* "change" in manual.

In the case of based on using the concept similarity clustering approach, the evaluations in CWN

and in *Xian Han* are presented in Table 6 below.

Table 6: Evaluations of the four target words in CWN and *Xian Han* based on the concept similarity clustering

Target Word	**CWN**			***Xian Han***		
	Sense	**Tagging**	**Recall**	**Sense**	**Tagging**	**Recall**
***Chi1* "eat"**	28	24	85.71%	8	7	87.5%
***Wan2* "play"**	9	9	100.00%	3	3	100%
***Huan4* "change"**	5	5	100.00%	3	3	100.00%
***Shao1* "burn"**	14	10	71.43%	8	4	50%
Average			**89.29%**			**84.38%**

From Table 6, we can not find 4 senses of *chi1* "eat" in CWN which are all metaphorical senses. For example, "比喻占便宜（flirt)" or "比喻有能力接受或完成（accept)". We either can not find 4 senses of shao1 "burn" in CWN which are all adjective senses. In the case of evaluations of in *Xian Han*, the results are the same as based on the character similarity clustering. Although the evaluations show that some senses also cannot be tagged in CWN and *Xian Han* based on using the concept similarity clustering approach, the important observation is that the recalls are better than based on using the character similarity clustering approach.

4 Conclusion

The aim of this sense prediction study is to explore all possible senses of lexical ambiguity in Mandarin Chinese by automatic prediction in machine programming. In this study, we use four corpora --- Chinese Gigaword Corpus, HowNet, Chinese Wordnet and *Xiandai Hanyu Cidian* as the database. The corpus-based and computational approach in this sense prediction study was aided by two main strategies: (1) character similarity clustering approach; and (2) concept similarity clustering approach. We are able to obtain higher accuracy rates and better performances using the concept similarity clustering approach in the sense prediction study. Regarded as the evaluation via Chinese Wordnet and *Xiandai Hanyu Cidian*, from these valuable evaluations of the character similarity clustering approach and the concept similarity clustering approach, we are able to demonstrate the viability of these two approaches as a superior resolution for this sense prediction study.

参 考 文 献

[1] Chen, Hao, Tingting He, Donghong Ji, and Changqin Quan. 2005. “An Unsupervised Approach to Chinese Word Sense Disambiguation Based on Hownet.” *Computational Linguistics and Chinese Language Processing.* 10:4, pp. 473–482.

[2] Dai, Liu-Ling, Bin Liu, Yuning Xia, and Shi-Kun Wu. 2008. *“Measuring Semantic Similarity between Words Using HowNet.”* International Conference on Computer Science and Information Technology, pp. 601–5.

[3] Fujii, Hideo and Croft, W. Bruce (1993): A Comparison of Indexing Techniques for Japanese Text Retrieval. In: Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 1993. pp. 237-246.

[4] Huang, Chu-Ren, Elanna I. J. Tseng, Dylan B. S. Tsai, and Brian Murphy. 2003. “Cross-lingual Portability of Semantic Relations: Bootstrapping Chinese WordNet with English WordNet Relations.” *Languages and Linguistics.* 4.3: 509–532.

[5] Li, Wanyin, Qin Lu, and Ruifeng Xu. 2005. “*Similarity Based Chinese Synonym Collocation Extraction.*” Computational Linguistics and Chinese Language Processing. 10:1, pp. 123–44.

[6] Liu, Qun and Su-Jian Li. 2002. *“The Word Similarity Calculation on <<HowNet>>.”* Proceedings of the 3rd Conference on Chinese lexicography, Taipei.

[7] Moldovan, Dan and Adrian Novischi. 2004. “Word sense disambiguation of WordNet glosses.” *Computer Speech and Language*, 18: 301–17.

[8] Peng, Jin, Xu Sun, Yunfang Wu, and Shiwen Yu. 2007. “Word Clustering for Collocation-Based Word Sense Disambiguation.” The Eighth International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2007), LNCS 4394, pp. 267–274.

[9] Xue, Nianwen Jinying Chen, and Martha Palmer. 2006. “Aligning Features with Sense Distinction Dimensions.” Proceedings of the COLING/ACL Main Conference Poster Sessions, pp. 921–928. Sydney, July 2006.