

Using Corpus-based Linguistic Approaches in Sense Prediction Study*

Jia-Fei Hong^a, Sue-Jin Ker^b, Chu-Ren Huang^c and Kathleen Ahrens^d

^aInstitute of Linguistics, Academia Sinica, Taiwan,
No. 128, Section 2, Academia Road 115, Taipei, Taiwan
jiafei@gate.sinica.edu.tw

^bDepartment of Computer Science and Information Management, Soochow University, Taiwan,
No.56, Section 1, Guiyang Street 100, Taipei, Taiwan
ksj@cis.scu.edu.tw

^cFaculty of Humanities, The Hong Kong Polytechnic University, Hong Kong
The Hong Kong Polytechnic University, Hong Hum, Hong Kong
churenhuang@gmail.com

^dLanguage Centre, Hong Kong Baptist University, Hong Kong
Hong Kong Baptist University, Kowloon Tong, Kowloon, Hong Kong
kathleenahrens@yahoo.com

Abstract. In this study, we propose to use two corpus-based 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 and *Xiandai Hanyu Cidian*. Using these corpora, we will determine the 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, concept similarity clustering, 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 for 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.

If a word has more than two senses at the same time, then it is usually called a lexically ambiguous word, for example, *bank*. We need to divide its correct senses and assign its appropriate senses to the different contexts. However, “lexical ambiguity” is presented in several different related researches to present and refer the same target. In WordNet, the definition of a lexically ambiguous word is the ambiguity of an individual word or phrase that can be used (in different contexts) to express two or more different meanings. WordNet researchers also regard polysemy and lexical ambiguity as synonym. In the case of the

* Copyright 2010 by Jia-Fei Hong, Sue-Jin Ker, Chu-Ren Huang and Kathleen Ahrens

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.

Although a suite of heuristical methods are presented for word sense disambiguation of Chinese Wordnet glosses, unfortunately 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 relevant 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.

In this study, we first review some related previous studies of sense prediction and lexical ambiguity resolution. Second, we point out lexical ambiguity determination, hypotheses, research questions, and the goal of this study. Third, we will introduce many important corpora and tools, such as Chinese Gigaword Corpus, HowNet, Chinese Wordnet and *Xiandai Hanyu*. Further, data collection, two main research approaches and evaluation will be shown and explained. Finally, we will show some results of combining, comparing, and discussing for these two approaches in this study.

2 Research Question

When we seek to 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. Lexical ambiguity means some words have multiple meanings or senses (Moldovan and Novischi, 2004). Therefore, the first hypothesis is that words with similar morpheme-character components and concept elements are similar in sense. We will follow Fujii and Croft (1993) to observe character similarity and refer to Li et al. (2003) and Dai et al. (2008) to explore concept similarity via HowNet. Peng et al. (2007) mentioned that a corpus was divided into five equal parts which one part was used as the test corpus and the collocation list was constructed from the other four parts of corpus. In this study, the second hypothesis is that different corpora with particular functions which provide different lexical knowledge bases. we will use Chinese Gigaword Corpus to select related collocations for the four target words; we will use HowNet to assign all possible concepts to ambiguous senses of the four target words; we will use Chinese Wordnet to estimate the evaluations for the four target words and we will also use *Xian Han* to estimate the evaluations for the four target words. 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 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 Chinese Wordnet (CWN) and Xiandai Hanyu Cidian (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 (Traditional 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 in order to predict their different senses for these four target words in the concept similarity clustering analysis of the corpus-based and computational 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 is an on-line common-sense knowledge base that reveals the inter-conceptual and inter-attribute relations of concepts as connoted in the Chinese lexicon and that of their English equivalents. HowNet includes an abundance of both semantic and world knowledge and thus is an important resource for NLP and knowledge mining (Dong and Dong 2006).

For the evaluations of the four target words, we will use Chinese Wordnet (CWN) and *Xiandai Hanyu*. The architecture of CWN follows the standard established by Princeton's WordNet (WN, Fellbaum 1998), which has two unique design features. First, it aims to maintain the balance between the universality of cross-lingual synset-based sense mapping and the felicity of language-specific lexicalization of concepts. Second, it aims to represent sense at the level of lexical conventionalization, as well as meaning facets at the level of conceptual specification (Ahrens et al. 1998). In addition, we then take *XianDai HanYu CiDian (Xia Han*, the fifth version, 2005) to evaluate them in the same time.

In order to gain these related collocations, from Taiwan's Central News Agency Gigaword Corpus, we use three 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 <u>吃鱼</u> 也有益健康。 | 鱼{Na} |
| 2 | 一种绝不含油脂的减肥食谱--盐包鸡，又称减肥鸡，并保证有 <u>吃五味盐酥鸡</u> 的滋味与口感。 | 盐酥鸡{Na} |
| 3 | 巴拿马人自己不 <u>吃猪内脏与猪脚筋</u> ，台湾二千三百万人口，每人每年平均消费四十点五公斤的猪肉，引起他们莫大的兴趣。 | 猪脚{Na} |
| 4 | 从 <u>玩彩色木棒或积木块的游戏</u> 中，能轻易学到像是长、高、形状、表面、尺寸。 | 游戏{Na} |
| 5 | 喝完汤后先吃蔬菜，尤以叶菜类及瓜类热量最低， <u>蔬菜</u> 尽量以凉拌或生 <u>吃</u> ，不加油更佳。 | 蔬菜{Na} |

Following these five different collocation selection criteria, there are 29,421 sentences for the collocation words of *chi1* "eat"; 8,833 sentences for the collocation words of *wan2* "play"; 19,394 sentences for the collocation words of *huan4* "change"; and 4,668 sentences for the collocation words of *shao1* "burn". From these sentences, there are 3,961 collocation words for

chi1 “eat”; 2,086 collocation words for *wan2* “play”; 3,003 collocation words for *huan4* “change”; and 1,565 collocation words for *shao1* “burn”. This empirical data can then be used to process the character similarity clustering analysis and the concept similarity clustering analysis using a corpus-based and computational approach; moreover, it can predict all of their possible senses.

4 Analysis

4.1 Character similarity clustering analysis

Following Fujii and Croft’s study (1993), we will use character similarity to cluster relative 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:

Formula 1: Character similarity comparison between words

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

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.

Formula 2: Group similarity comparison between words

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

In Formula 2, *x* represents one undefined word, while *y* represents each word of *Y*, where *Y* indicates a particular cluster. To determine which words belong in which clusters, first, one undefined word (*x*) must be compared with another word (*y*), then their average similarity must be calculated in order to gain the maximum similarity, and, finally, this undefined word (*x*) is placed into a particular cluster (*Y*). After comparing the cluster similarities, 败绩 (*bai4 ji1* “defeat”) and 败仗 (*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.

Formula 3: Average similarity of two different clusters

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

In Formula 3, *clu₁* and *clu₂*, respectively, represent different clusters; *s* and *t*, respectively, are the word members of *clu₁* and *clu₂*; and $|clu_1|$ and $|clu_2|$, individually, represent the number of clusters of *clu₁* and *clu₂*. Hence, 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 these words have analyzed in Chinese Wordnet already. 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, collocations with frequencies that are less or equal two (\leq 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 their accuracy, the testing cluster sizes will be 1, 1.5, and 2 times that of the senses. In other words, when examining the accuracy of the four target words in the character similarity clustering analysis, we will select the top 30 clusters, 45 clusters, and 60 clusters of the 100 clusters, 200 clusters and 300 clusters for *chi1* “eat”, respectively; the top 10 clusters, 15 clusters, and 20 clusters of the 90 clusters, 180 clusters, and 270 clusters for *wan2* “play”, respectively; the top 6 clusters, 9 clusters, and 12 clusters of the 70 clusters, 140 clusters, and 210 clusters for *huan4* “change”, respectively; and the top 15 clusters, 23 clusters, and 30 clusters of the 60 clusters, 120 clusters, and 180 clusters of *shao1* “burn”, respectively. 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.

4.2 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, 山藥 (*shan1 yao4* “Chinese yam”) and 藥 (*yao4* “medicine”) are in the same cluster. 漢堡肉 (*han4 bao3 rou4* “hamburger meat”) is categorized into 漢堡 (*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. We will extract the sememes of the

concept for each collocation word by HowNet. 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.

Formula 4: Similarity between sememes

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

Among them, $d(S1)$ and $d(S2)$ represent the level of sememe $S1$, and $S2$, separately, in the semantic concept tree, while $\text{dis}(S1, S2)$ represents the distance of sememe $S1$ and $S2$ in the semantic concept tree.

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.

Formula 5: Similarity between concepts

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

In Formula 5, pm and pn represent the primary sememes of concept m and concept n , separately. We then calculate similarity, which is $\text{sim_seme}(pm, pn)$, between the main sememes of two concepts. Then we gain the final average similarity via Formula 5 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. Taking the same method as the character similarity clustering analysis, we also randomly select some clusters as my 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 precisions are all over 84% and the average precision is 85.90%.

Table 4: Average precision for four target words

| Target Word | Accuracy Rate |
|-----------------------|---------------|
| <i>Chi1</i> “eat” | 85.59% |
| <i>Wan2</i> “play” | 87.21% |
| <i>Huan4</i> “change” | 85.98% |
| <i>Shao1</i> “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.

4.3 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 assigned appropriate senses. In CWN, there are 28 senses for *chi1* “eat”, 9 senses for *wan2* “play”, 5 senses for *huan4* “change”, and 14 senses for *shao1* “burn”, and in *Xian Han*, there are 8 senses for *chi1* “eat”, 3 senses for *wan2* “play”, 3 senses for *huan4* “change”, and 8 senses for *shao1* “burn”. However, we only focus on the transitive verbs in this sense prediction; we need to remove the noun usage sense and non-transitive verb usage senses in CWN and *Xian Han*. In addition, we only concentrate on the modern Chinese; we also need to remove early period vernacular usage senses. For the evaluations in this study, the number of senses in CWN and *Xian Han* are presented in Table 5 below.

Table 5: Number of senses in CWN and *Xian Han*

| Target Word | Chinese Wordnet | Xiandai Hanyu Cidian |
|-----------------------|-----------------|----------------------|
| <i>Chi1</i> “eat” | 28 | 7 |
| <i>Wan2</i> “play” | 9 | 3 |
| <i>Huan4</i> “change” | 5 | 3 |
| <i>Shao1</i> “burn” | 13 | 5 |

Following the principle of calculating the accuracy rates of *chi1* “eat”, *wan2* “play”, *huan4* “change”, and *shao1* “burn” in the character similarity clustering analysis, we selected the top-down 2-times number of clusters as my testing data. In addition, we also selected the bottom-up 1-time number of clusters as my other testing data in order to examine whether we could find other senses that do not appear in CWN or *Xian Han*, or whether we could find new appearances of the four target words. The distribution of the number of clusters using the character similarity clustering approach is shown in Table 6 below.

Table 6: Number of clusters for evaluation using the character similarity clustering approach

| Target Word | Total Cluster | Testing Cluster | |
|-----------------------|---------------|-----------------|-----------|
| | | Top-down | Bottom-up |
| <i>Chi1</i> “eat” | 200 | 60 | 30 |
| <i>Wan2</i> “play” | 180 | 20 | 10 |
| <i>Huan4</i> “change” | 140 | 12 | 6 |
| <i>Shao1</i> “burn” | 120 | 30 | 15 |

Based on the character similarity clustering analysis, the distributions of the sense prediction evaluations for the four target words in CWN and *Xian Han* are shown in Table 7 below.

Table 7: Evaluations in CWN and *Xian Han* based on the character similarity clustering

| Target Word | CWN | | | <i>Xian Han</i> | | |
|-----------------------|-------|---------|---------------|-----------------|---------|---------------|
| | Sense | Tagging | Recall | Sense | Tagging | Recall |
| <i>Chi1</i> “eat” | 28 | 22 | 78.57% | 7 | 7 | 100.00% |
| <i>Wan2</i> “play” | 9 | 8 | 88.89% | 3 | 3 | 100.00% |
| <i>Huan4</i> “change” | 5 | 5 | 100.00% | 3 | 3 | 100.00% |
| <i>Shao1</i> “burn” | 13 | 8 | 61.54% | 5 | 4 | 80.00% |
| Average | | | 82.25% | | | 95.00% |

From Table 7, the evaluations show that some senses cannot be tagged in CWN and *Xian Han* based on using the character similarity clustering approach.

In the case of based on using the concept similarity clustering approach, the calculations and recalls in CWN and in *Xian Han* are presented in Table 8 below.

Table 8: Recall of the four target words in CWN based on the concept similarity clustering

| Target Word | CWN | | | <i>Xian Han</i> | | |
|-----------------------|-------|---------|---------------|-----------------|---------|---------------|
| | Sense | Tagging | Recall | Sense | Tagging | Recall |
| <i>Chi1</i> “eat” | 28 | 24 | 85.71% | 7 | 7 | 100.00% |
| <i>Wan2</i> “play” | 9 | 9 | 100.00% | 3 | 3 | 100.00% |
| <i>Huan4</i> “change” | 5 | 5 | 100.00% | 3 | 3 | 100.00% |
| <i>Shao1</i> “burn” | 13 | 10 | 76.92% | 5 | 4 | 80.00% |
| Average | | | 90.66% | | | 95.00% |

From Table 8, 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 in *Xian Han* are better than in CWN whether based on using the character similarity clustering approach or whether based on using the concept similarity clustering approach.

5 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. We use corpus-based analysis and evaluate their performances. 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. In the character similarity clustering approach, character similarity was compared between words and then grouped according to their similar morphemes. In the case of the concept similarity clustering approach, we mapped all possible concepts for all the collocation words of the four target words using HowNet; hence, two important strategies emerged: (1) similarity between sememes; and (2) similarity between concepts. We then clustered some collocation words into the same cluster based on their concepts in order to predict all possible concepts. 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 model for this sense prediction study.

References

- Ahrens, Kathleen. 1998. “Lexical Ambiguity Resolution: Languages, Tasks and Timing.” *In Sentence Processing: A Cross-linguistic Perspective*. (Ed.) Dieter Hillert. Academic Press, pp.11–31.
- Canas, Alberto J., Alejandro Valerio, Juan Lalinde-Pulido, Marco Carvalho, and Marco Arguedas. 2003. “Using WordNet for Word Sense Disambiguation to Support Concept Map Construction.” Paper presented at SPIRE 2003—10th International Symposium on String Processing and Information Retrieval, Oct. 2003, Manaus, Brazil.
- 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.

- Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.
- 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.
- 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.
- Ganesh, Ramakrishnan, B. P. Prithviraj, A. Deepa, Pushpak Bhattacharyya, and Soumen Chakrabarti. "Soft Word Sense Disambiguation." 2004. *GWC 2004, Proceedings*, pp. 291–298.
- Ker, Sue-Jin and Jen-Nan Chen. 2004. "Adaptive Word Sense Tagging on Chinese Corpus." *PACLIC 18*, Dec. 8–10, 2004, Waseda University, Tokyo, pp. 267–273.
- 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.
- Liu, Qun and Su-Jian Li. 2002. "The Word Similarity Calculation on <<HowNet>>." *Proceedings of the 3rd Conference on Chinese lexicography*, Taipei.
- Martinez, David, Eneko Agirre, and Xinglong Wang. 2006. "Word Relatives in Context for Word Sense Disambiguation." *Proceedings of the 2006 Australasian Language Technology Workshop (ALTW2006)*, pp. 42–50.
- Moldovan, Dan and Adrian Novischi. 2004. "Word sense disambiguation of WordNet glosses." *Computer Speech and Language*, 18: 301–17.
- 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.
- Ravin, Yael, and Claudia Leacock. 2000. *Polysemy: An overview. Polysemy: Theoretical and computational approaches*, ed. by Yael Ravin and Claudia Leacock, 1-29. New York: Oxford University Press.
- Resnik, Philip and David Yarowsky. 2000. "Distinguishing Systems and Distinguishing Senses: New Evaluation Methods for Word Sense Disambiguation." *Natural Language Engineering* 5 (3): 113–33. Printed in the United Kingdom. Cambridge University Press.
- Veronis, Jean and Nancy M. Ide. 1990. "Word sense disambiguation with very large neural networks extracted from machine readable dictionaries," *Proceedings of the 13th Conference on Computational Linguistics*, pp. 389–94, August 20–25, 1990, Helsinki, Finland.
- 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.
- Zhang, Yuntao, Ling Gong, and Yongcheng Wang. 2005. "Chinese Word Sense Disambiguation Using HowNet." L. Wang, K. Chen, and Y.S. Ong (Eds.): *ICNC 2005*, LNCS 3610, pp. 925–32.