KNOWLEDGE-BASED SENSE PRUNING USING THE HOWNET: AN ALTERNATIVE TO WORD SENSE DISAMBIGUATION

GAN Kok-Wee[1], *WANG Chi-Yung*[2] *and Brian MAK*[3]

wcyung@cs.ust.hk[2], mak@cs.ust.hk[3]
Department of Computer Science
The Hong Kong University of Science & Technology

gankw@dbi.org.hk[1]
The Dharmasthiti College
of Cultural Studies

ABSTRACT

Word sense disambiguation (WSD) is one of the basic problems in natural language processing. Traditional WSD methods provide only one meaning for each word in a passage. However, we believe that textual information alone may not be sufficient to determine the exact meaning of each word which may better be resolved when higher-level knowledge becomes available. In this paper, we propose an alternative to WSD that we call "*sense pruning*". The objective now is to reduce the number of plausible meanings of a word as much as possible so as to reduce the amount of work in later processing. Sense pruning is guided by information derived from HowNet —– a recently developed knowledge base.

Two criteria were used for the evaluation: recall rate and complexity reduction (which is the reduction in the number of possible meanings of a sentence). Effect of the length of the analytical window was studied. For a corpus of 103 Chinese passages from Sinica, Taiwan, with an analytical window of nine words, we obtained a recall rate of 94.14% and reduced the number of possible sentence meanings by 65.3%.

1. INTRODUCTION

In *re-constructive text understanding* [4], raw texts (without any tagged information) undergo five NLP steps before they are interpreted. Gan and Wong [8] defined five NLP steps:

- sentence breaking;
- concept group extraction;
- sense pruning;
- message structure identification; and
- event relation and role-shifting.

The idea is to progressively apply more knowledge to narrow down the plausible meanings of a given text. It is important to retain all possible answers at each NLP step when there is insufficient knowledge to prove them irrelevant. Thus, sense pruning is employed in the third stage where traditionally word sense disambiguation would have been used because the latter will otherwise undesirably produce only one single meaning for later processing. In sense pruning, unlikely senses of each word are pruned away, but usually more than one sense for each word is retained.

In this paper, sense pruning makes use of information derived from a recently developed knowledge resource called HowNet [2, 3, 4, 5]. Advantages of using HowNet owe to its richness and language-independence. A score is computed for each sense of a word by taking into account all the senses of surrounding words within an analytical window. Those senses that score below a threshold will be pruned away. We evaluated our new sense-pruning algorithm on a corpus from Sinica, Taiwan, consisting of 103 news articles on crime, and compared the results with those obtained by a traditional WSD algorithm. This paper is organized as follows: we first introduce the HowNet system in Section 2, and describe some related works using HowNet in Section 3. Section 4 describes how we made use of HowNet to perform sense pruning. Experimental evaluation is shown in Section 5 which is followed by our conclusions in Section 6.

2. THE HowNet KNOWLEDGE SYSTEM

The HowNet knowledge system was developed by Dong [2, 3,4,5] over the last decade. Its latest version was released online in October 2000. For detailed information about the system, the reader is referred to its website at http://www.keenage.com. Below we will give a short description of the system.

2.1. HowNet

HowNet is an online common-sense knowledge base that captures inter-conceptual relations and inter-attribute relations of concepts. As a knowledge base, HowNet is language-independent. In HowNet, the smallest basic semantic unit is called *sememe*. At present, Dong has identified 1,503 sememes, which are used to describe all concepts in the HowNet Knowledge Dictionary. Associated with each sememe is a set of attributes or dynamic roles. Dong believes that all matters, physical or non-physical, keeps changing continually. Their changes are reflected in a change of their internal state, which in turn, is manifested in a change of some of their attribute values.

The HowNet knowledge system consists of the HowNet Knowledge Dictionary and the HowNet Management System.

```
NO.=056352
W_C=知網
G_C=N
E_C=
DEF=software|軟件, #knowledge|知識
```

Figure 1 An example entry in the Knowledge Dictionary

2.1.1. HowNet Knowledge Dictionary

The HowNet Knowledge Dictionary is bilingual in Chinese and English. While a regular dictionary is for words, the HowNet Knowledge Dictionary is for concepts. The latest version (HowNet 2000) covers over 110,000 concepts in the Dictionary. For example, the entry for the concept "HowNet" (知網) in the Dictionary is shown in Figure 1. The entry tells us that the concept "知網" is the 56352-th entry in the Dictionary. Its written form is given by: W_C=知網; it is used as a noun (G_C=N); no example of its usage is given and thus the field E_C is empty; and finally its definition is given by the field DEF and is a software related to the sememe of "knowledge". The first definition in the DEF field is called the main feature of the concept, and the remaining definitions are its secondary features. There are two major classes of main features: event (事件) and entity (實體). All definitions are expressed in terms of the basic sememes whose details are depicted and organized in a set of supplementary documents in the HowNet Management System.

2.1.2. Documents of the HowNet Management System

The HowNet Management System organizes all sememes in ten documents: (1) structure of main features, (2) list of attributes, (3) list of attribute values, (4) list of quantities, (5) list of quantity values, (6) list of secondary features, (7) list of event roles and attributes, (8) list of syntaxes, (9) list of antonyms, and (10) list of converses. All the ten documents are designed in a simple list format with the exception of the structure of main features, which is organized as a tree. Together with the Knowledge Dictionary, one may use these documents to query the attributes or dynamic roles of a concept and their values.

2.2. Information Structure

Information Structure is an extension of HowNet. Since the same concept may be expressed syntactically differently in different languages, Information Structure tries to capture the syntactic patterns in a language. Thus, it is language dependent. Currently it is only available for the Chinese language. The Chinese HowNet Information Structure has defined more than 279 types of patterns. It is further classified into 45 relationships, such as "composite" (合成), "modifier" (修飾), "coordinate" (並列), etc. There is at least one type of pattern under each relationship. A pattern is a pair of sememes in the following format:

(sememe) [relationship] ← (sememe) .

3. RELATED WORKS

Since HowNet is a very new knowledge resource, there are few applications to date. Yang, Zhang, and Zhang [6] used HowNet to perform WSD using a statistical approach. Co-occurrence frequencies between any pair of sememes were computed from a corpus consisting of news articles from Peoples' Daily with 10,000 characters, resulting in 709,496 entries. The accuracy of the system is around 75%.

4. SENSE PRUNING

In our proposed sense-pruning algorithm, the senses of a word token are pruned according to four types of information derived from HowNet under the context of its surrounding words, which make up the analytical window. For each sense of the given word, we compute a score for the sense using the following four types of information. Those senses that score below a threshold will be pruned.

4.1.1. Sememe Co-occurrences

Scoring by sememe co-occurrences may be explained using pseudo-codes as in Algorithm 1.

Algorithm 1 Scoring with sememe co-occurrences

```
foreach sense of a given word
{
    score = 0;
    get the set of sememes X that describes the sense;

    foreach sense of another word in the analytical window
    {
        get the set of sememes Y that describes the sense;
        matches = number of common sememes in X and Y;
        score += matches;
    }
}
```

Let us illustrate with an example. If we have "suffer" and "patient" within an analytical window, we first look up their senses from the Knowledge Dictionary and get Table 1 and 2.

Table 1 Definitions of "suffer"

Concept	Definition
suffer 患	1. sufferFrom|罹患, medical|醫
	2. emotion|情感, undesired|莠, #sad|憂愁
	3. phenomena|現象, undesired|莠, #unfortunate|不幸
	4. phenomena|現象, undesired|莠, hardship|艱, #unfortunate|不幸
	5. sad|憂愁

Table 2 Definition of “patient”

Concept	Definition
patient 病人	human|人, *sufferFrom|罹患*, $cure|醫治, *#medical|醫*, undesired|莠

Table 3 Definitions of “little” and “boy”

Concept	Definition
little 小	1. aValue|屬性值, *age|年齡*, young|幼
	2. aValue|屬性值, duration|久暫, ?timeShort|暫
	3. aValue|屬性值, importance|主次, secondary|次
	4. aValue|屬性值, size|尺寸, small|小
boy 朋友	*human*|人, friend|友

From the two tables, we find five senses for “suffer” and one sense for “patient”. The numbers of sememe co-occurrences between each sense of “suffer” and “patient” are: 2, 1, 1, 1, 0 respectively, while that between “patient” and “suffer” is 5.

Notice that sometimes we have to make use of the hypernymy-hyponymy relationship and the list of main features to obtain the sememe co-occurrences indirectly.

4.1.2. Information Structure

Similarly to Algorithm 1, we examine each combination of a sememe from the set X and a sememe from the set Y, and if they represent a pattern in the HowNet Information Structure, we increment the score of each respective sense of the two words by one.

4.1.3. Attributes of an Object

Sememe co-occurrences may also be searched through the list of attributes of an object belonging to the entity (main feature) class. For example, in the phrase “little boy”, if one looks at their definitions in the Knowledge Dictionary (Table 3), one will not find any common sememes. However, from the entity class tree, a hypernym of the main feature “human” in the definition of “boy” is “animate|生物”, and according to the document of the list of attributes, it has the “age” (年齡) attribute. Thus, the first definitions of “little” and “boy” have one common sememe (“age”) and both of these senses get an additional score.

4.1.4. Special Patterns for Functional Words

HowNet is comparatively weak in its treatment with functional words. It only provides the functional definitions (using curly brackets) of functional words. We create new special patterns for functional words with the following format:

functional-word {definition} <related sememe|feature> .

For example, a pattern for the functional word “at” is:

at (在) {location} <place> .

When one of any two words in an analytical window is a functional word, these special patterns are checked and if there is a match, their corresponding senses are scored.

Table 4 Results when a complete sentence is used as the analytical window

Recall Rate	Complexity Reduction
97.13%	47.63%

5. EVALUATION

5.1. Corpus

We evaluated our new sense-pruning algorithm using 103 newspaper passages in crime domain from the Sinica corpus, version 3.0 of Academia Sinica [1]. The articles consists of ~30,000 Chinese words. On average, there are about 234 word tokens in a passage, 45 words in a complete sentence, and 7.7 words per incomplete sentence.

5.2. Methodology

All the passages were re-segmented into word tokens that represent legal concepts in the HowNet Knowledge Dictionary. For each word token, an analytical window with a length of N words (where N was varied) was left aligned with it, and each of its sense was scored according to the four types of information explained in Section 4. In this paper, scores from each type of information were weighted equally, and those senses with zero score were pruned. In the rare case when all senses of a word token did not score, we simply kept all its senses to avoid having a sentence with no senses.

5.3. Criteria

Two measures were used to gauge the performance of our sense-pruning algorithm.

1. **Recall Rate:** the percentage of word tokens whose correct sense is not pruned.
2. **Complexity Reduction:** the average ratio of the number of possible meanings of a sentence after sense pruning to the total number of possible meanings of the sentence before sense pruning.

5.4. Results

5.4.1. Experiment 1: A Complete Sentence Used as the Analytical Window

We first tried to use a complete sentence as the analytical window. Thus, the window length is actually changing from sentence to sentence. The hypothesis is that all words in a sentence help determine the sense of any word in the sentence; but the words in a sentence have no effect on the sense of a word in another sentence. The results are shown in Table 4. The high recall rate shows that the information provided by the HowNet knowledge system is very powerful for NLP research, especially given that HowNet is domain-independent. However, there still is room for improvement. For instance, we find that some definitions of functional words pertaining to sememes that are not well defined. One example is the list of dynamic roles of sememes in the event (事件) class.

5.4.2. *Experiment 2: Effect of Window Length*

In this experiment, we would like to investigate the effect of a distant word on the sense of another word in the same sentence by varying the length of the analytical window. The results are plotted in Figure 2. As expected, the recall rate improves with larger analytical window when more information are provided to resolve for the right sense of a word; meanwhile, the complexity reduction decreases accordingly as more combination of word senses are possible with larger window. On the other hand, the recall rate only increases slowly at a rate of ~1% for every two additional words in the analytical window. Thus, the immediate neighboring word is still the most important but words at a distance as far as nine words away still have effects in determining the right sense of a word!

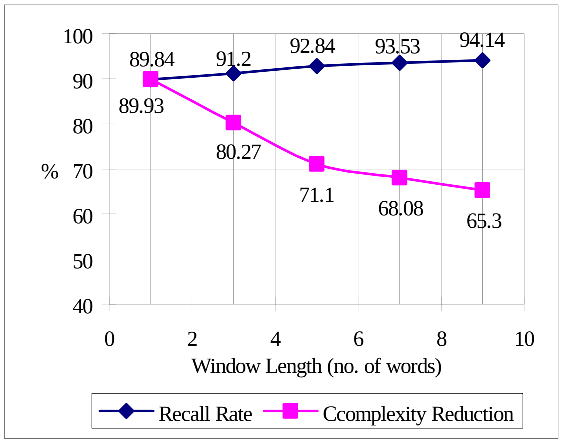

Figure 2 Effect of window length

5.4.3. *Experiment 3: Comparison with a WSD Baseline*

We implemented a simple statistical word sense disambiguation method as a baseline for comparison. As a first step, we had to create the sememe co-occurrence frequency table. We divided our evaluation corpus of 103 passages into five disjoint sets (of roughly the same size) and performed a 5-fold cross-validation. That is, WSD was performed five times, each time with a different set held out for testing while the remaining four sets were used for training the co-occurrence frequency table. Notice that unlike sense pruning, WSD picks only the sense of a word with maximum frequency. The mean recall rate of the 5-fold cross-validation is 85.07%, which is significantly lower than the result of sense pruning with a window length of 1 (which is 89.84%). If one further takes into account that the co-occurrence frequencies in the baseline WSD experiment were trained on texts similar to the testing texts while the HowNet knowledge system is domain-independent, one should be convinced that HowNet indeed is a very powerful knowledge base. (On the other hand, WSD has the advantage of attaining a complexity reduction of almost 100% since there is only one sense for the whole sentence.)

6. CONCLUSION

In this paper, we have two contributions: (1) we investigated the performance of a sense pruning algorithm which is a crucial step in re-constructive text understanding to replace traditional word sense disambiguation; and, (2) we pioneered to employ the domain-independent HowNet knowledge system to perform sense pruning. Our results show that sense pruning compares favourably with word sense disambiguation in its high recall rates even without domain-specific knowledge. However, there is still a lot to be done to further reduce the overall sentence complexity in order to alleviate the workload for the following steps in re-constructive text understanding. Some parameters in our sense-pruning algorithm, such as the weightings of the four types of HowNet information and the pruning threshold, also need further investigation. The high recall rate is evidence that the HowNet knowledge base has a broad coverage. On the other hand, we also point out some areas of concern in the HowNet systems: it should further perfect the definitions of functional words and sememes related to the event class.

7. ACKNOWLEGEMENTS

This work is supported by the Sino Software Research Institute under the grant number of SSRI99/00.EG06.

8. REFERENCES

[1] CKIP, "*The Content and Illustration of Sinica Corpus of Academia Sinica*," Technical Report no. 95-02 (*中央研究院平衡語料庫的內容與說明, 技術報告 95-02*). Institute of Information Science, Academia Sinica, 1995.

[2] Dong, Zhendong, "*Knowledge Description: What, How and Who?*" In Proceedings of International Symposium on Electronic Dictionary, Tokyo, Japan, 1998.

[3] Dong, Zhendong, "*HowNet*." http://www.keenage.com, 1999.

[4] Dong, Zhendong, "*Bigger Context and Better Understanding – Expectation on Future MT Technology*." In Proceedings of International Conference on Machine Translation and Computer Language Information Processing, 26-28 June, 1999, Beijing, China, pp. 17-25.

[5] Dong Zhendong, "*The Pattern of Chinese Information Structure*" (*中文信息結構模式*), ms.

[6] Erhong Yang, Guoqing Zhang and Yongkui Zhang, "*The Research of Word Sense Disambiguation Method Based on Co-occurrence Frequency of HowNet*." In Proceedings of the 2nd Chinese Language Processing Workshop, Association for Computational Linguistics 2000 Conference, October 2000, Hong Kong, pp. 60-65.

[7] Gan, Kok-Wee and Tham, Wai-Mun, "*General Knowledge Annotation Based on HowNet*" (*基於知網的常識知識標註*). Computational Linguistics and Chinese Language Processing, vol. 4, 1999, pp. 39-86.

[8] Gan, Kok-Wee and Wong, Ping-Wai, "*Annotating Information Structures in Chinese Text using HowNet*." In Proceedings of the 2nd Chinese Language Processing Workshop, Association for Computational Linguistics 2000 Conference, October 2000, Hong Kong, pp. 85-92.