Word Sense Disambiguation and Sense-Based NV Event Frame Identifier

Jia-Lin Tsai, Wen-Lian Hsu and Jeng-Woei Su
Institute of Information Science, Academia Sinica
Nankang, Taipei, Taiwan, R.O.C.
{tsaijl,hsu,raycs}@iis.sinica.edu.tw

Abstract

Word sense is ambiguous in natural language processing (NLP). This phenomenon is particularly keen in cases involving noun-verb (NV) word-pairs. In Chinese processing, there is an additional difficulty in word segmentation. This paper describes a sense-based noun-verb event frame (NVEF) identifier that can be used to disambiguate word sense in Chinese sentences effectively. A knowledge representation system (the NVEF-KR tree) for the NVEF sense-pair identifier is also proposed. We use the word sense in Hownet, which is a Chinese-English bilingual knowledge-base dictionary.

Our experiment shows that the NVEF identifier is able to achieve 74.8% accuracy for the test sentences based only on NVEF sense-pair knowledge. By applying the techniques of longest syllabic NVEF-word-pair first and exclusion word checking, the sense accuracy for the same test sentences can be further improved to 93.7%. There are four major reasons for the incorrect cases: (1) lack of a bottom-up tagger, (2) lack of non-NVEF knowledge, (3) inadequate word segmentation, and (4) lack of a multi-NVEF analyzer. If these four problems can be resolved, the accuracy will be 98.9%.

The result of this study indicates that NVEF sense-pair knowledge is effective for word sense disambiguation and it is likely to be important for general NLP.

Keywords: word sense disambiguation, event frame, top-down identifier, Hownet

1. Introduction

Word sense disambiguation (WSD) has been a pervasive problem in natural language processing (NLP) since 1949 [1]. Word sense ambiguity (or lexical ambiguity), is generally classified into two types: syntactic and semantic ambiguity [2,3]. Syntactic ambiguity is caused by differences in syntactic category (e.g. “play” can occur as a *noun* or *verb*). Semantic ambiguity is caused by homonymy (e.g.

"bank" in "*to put money in a bank*", "*the bank of a river*") or polysemy (e.g. "face" in "*human face*", "*face of a clock*"). Although many approaches have been adopted to disambiguate word sense, algorithms for word sense determination are still not reliable [3,4]. Human beings usually can disambiguate word sense by using additional information from the speaker, the writer or the context. When out-of-context (or out-of-sentence) information are not symbolized and processed in computer, WSD either becomes very difficult or, sometimes, impossible. Therefore, it is crucial to investigate what kind of knowledge can be useful for WSD [3].

According to a study in cognitive science [5], people often disambiguate word sense with only a few other words in the context (frequently only one additional word). Thus, the relationships between a word and others can be effective information for resolving ambiguity. Furthermore, from [2,3,4], most ambiguities occur on nouns and verbs and object-event (i.e. noun-verb) distinction is a major ontological division for human [6]. However, no clear data has been collected to support these claims. These observations motivate us to demonstrate through an experiment, how noun-verb (NV) relationships can be used to disambiguate word sense in Chinese sentences.

In this paper we shall focus on word sense disambiguation involving NV word-pairs since these are most troublesome. Consider the following sentence, "這輛車行駛順暢(This car moves well)". In this sentence, we have two possible NV word-pairs, "車-行駛 (car, move)" and "車行-駛(auto-shop, move)". It is clear that the permissible NV word-pair is "車-行駛(car, move)". We shall call such a permissible NV word-pair a *NV-event frame* (*NVEF*). Using a collection of pre-learned NVEF, we can identify the NVEF word-pair, "車-行駛", from the sentence "這輛車行駛順暢". The word "車" in dictionary can have three possible senses: 'surname' (noun), 'car' (noun) and 'turn' (verb). To resolve this ambiguity, we can use the pre-defined sense of the NVEF, "車-行駛(car, move)", to determine that the correct sense of the Chinese word "車" is "car" in the above Chinese sentence.

In this paper, we shall illustrate that the knowledge of NVEF sense-pair (to be defined in Section 2) is effective in resolving word sense ambiguity. In the next section, we propose a NVEF sense-pair identifier, which is based on pre-stored knowledge of NVEF sense-pairs. We use this NVEF sense-pair identifier to identify NVEF word-pairs in the input sentence and to determine the corresponding word senses. In Section 3, we present and analyze the results of a WSD experiment for a set of test sentences using the NVEF sense-pair identifier. Finally, we give the conclusion and future direction in Section 4.

2. Development of a NVEF Sense-Pair Identifier

We use Hownet [7] as our system's Chinese machine-readable dictionary (MRD). Hownet is a Chinese-English bilingual knowledge-base dictionary, which provides the knowledge of Chinese lexicon, part-of-speech (POS) and word senses.

2.1 A NVEF Sense-Pair

The sense of a word is defined to be its DEF (concept definition) in Hownet. Table 1 lists three different senses of the Chinese word, "車(Che/car/turn)". In Hownet, the DEF of a word consists of its main feature and secondary features. For example, in the DEF, "character|文字,surname|姓,human|人,ProperName|專" of the word "車(Che)", the first item "character|文字" is the main feature, and the remaining three items "surname|姓", "human|人", and "ProperName|專" are its secondary features. The main feature in Hownet can inherit features in the hypernym-hyponym hierarchy. There are approximately 1,500 features in Hownet. Each of these features is called a *sememe*, which refers to a smallest semantic unit that cannot be further reduced.

Table 1. Three different senses of the Chinese word "車(Che/car/turn)"

C.Word [a]	E.Word [a]	Part-of-speech	Sense (i.e. DEF in Hownet)
車	Che	Noun	character|文字,surname|姓,human|人,ProperName|專
車	car	Noun	LandVehicle|車
車	turn	Verb	cut|切削

[a] C.Word refers to a Chinese word; E.Word refers to an English word

The Hownet dictionary used in this study contains 50,121 Chinese words, in which there are 29,719 nouns, 16,652 verbs and 16,242 senses (including 9,893 noun-senses and 4,440 verb-senses). Table 2 gives the statistics of the number of senses per Chinese word and the number of Chinese words per sense used in Hownet.

Table 2. Statistics of the number of senses per Chinese word and the number of Chinese words per sense used in Hownet

Item [a]	Total	Noun	Verb
Maximum number of senses per Chinese word	27	14	24
Mean number of senses per Chinese word	1.24	1.14	1.23
Maximum number of Chinese words per sense	374	372	129
Mean number of Chinese words per sense	3.8	3.0	4.6

[a] Similar statistics of WordNet can be found in [8] (WordNet is a trademark of Princeton University)

Now, take the NV word-pair "車-行駛 (car-move)" for example. According to the noun-sense of the Chinese word “車(Che/car/turn)” and verb-sense of the Chinese word “行駛(move)”, the only permissible NV sense-pair for the NV word-pair “車-行駛(car-move)” is “LandVehicle|車”-“VehicleGo|駛”. We call such a permissible NV sense-pair a *NVEF sense-pair* in this paper. Note that a NVEF sense-pair is a class that includes the permissible word-pair instance “車-行駛(car-move)”.

2.2 Knowledge Representation Tree of NVEF Sense-Pairs

A knowledge representation tree (KR-tree) of NVEF sense-pairs is shown in Fig.1. There are two types of nodes in the KR-tree, namely, ***function nodes*** and ***concept nodes***. Concept nodes refer to words and features in Hownet. Function nodes are used to define the relationships between their parent and children concept nodes. If a concept node A is the child of another concept node B, then A is a subclass of B. By this convention, we can omit the function node “subclass” (which should have existed) between A and B. We classify the noun-sense class (名詞詞義分類) into 15 subclasses according to their main features. They are “微生物(bacteria)”, “動物類(animal)”, “人物類(human)”, “植物類(plant)”, “人工物(artifact)”, “天然物(natural)”, “事件類(event)”, “精神類(mental)”, “現象類(phenomena)”, “物形類(shape)”, “地點類(place)”, “位置類(location)”, “時間類(time)”, “抽象類(abstract)” and “數量類(quantity)”. Appendix A gives a sample table of 15 main features of nouns in each noun-sense subclasses.

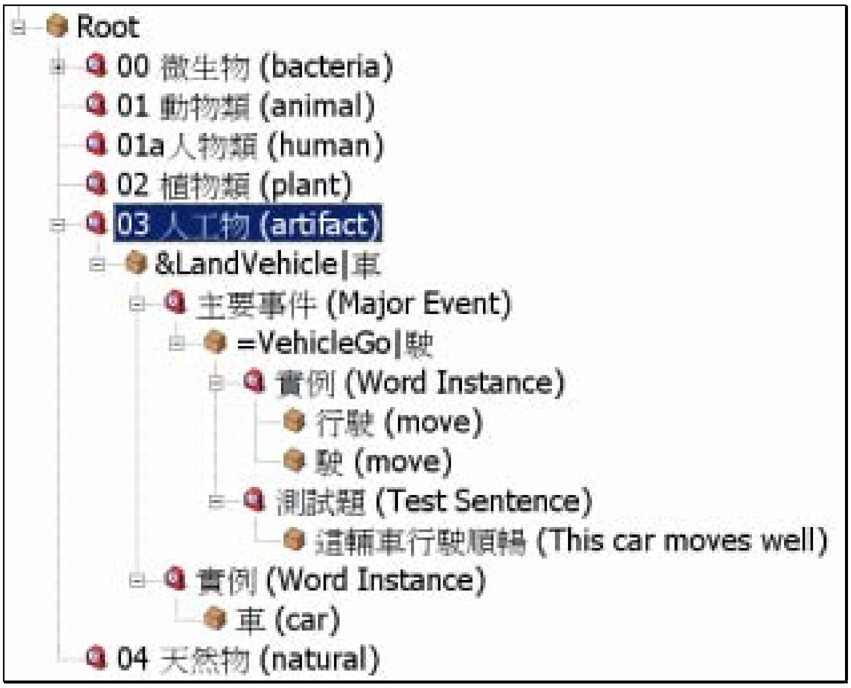

Figure 1. An illustration of the KR-tree using “人工物(artifact)” as an example noun-sense class (The English words in parentheses are there for explanatory purpose only).

There are three function nodes used in the KR-tree as described in Fig. 1:

(1) Major-Event (主要事件): The content of its parent node represents a noun-sense subclass and the content of its child node represents a verb-sense subclass. A noun-sense subclass and a verb-sense subclass linked by a Major-Event function node is a NVEF subclass sense-pair, such as "&LandVehicle|車" and "=VehcileGo|駛" in Fig. 1. To describe various relationships between noun-sense and verb-sense subclasses, we design the following three subclass sense-symbols, in which "=" means "*exact*", "&" means "*like*", "%" means "*inclusive*". An example of these symbols is given below.

Given three senses S_1, S_2 and S_3 defined by a main feature A and three secondary features B, C and D, let

S_1 = A, B, C, D

S_2 = A, B

S_3 = A, C, D

Then, we have that sense S_2 is in the "=A,B" *exact*-subclass; senses S_1 and S_2 are in the "&A,B" *like*-subclass; and senses S_1 S_2, and S_3 are in the "%A" *inclusive*-subclass.

(2) Word-Instance (實例): The content of its children are the words belonging to the sense subclass of its parent node. These words are self-learned by the NVEF identifier according to the sentences under the Test-Sentence nodes.

(3) Test-Sentence (測試題): The content of its children is several selected test sentences in support of its corresponding NVEF subclass sense-pair.

2.3 Generation of NVEF Sense-Pairs

To speedup the creation of the KR-tree, an example-based algorithm is proposed to generate the KR-tree semi-automatically. This algorithm is described below.

Step 1. Select a noun-sense such as "disease|疾病" in Hownet.

Step 2. Collect all Chinese polysyllabic words of the selected noun-sense (The monosyllabic words are not considered at this stage).

Step 3. Select those Chinese un-segmented sentences that include at least one word collected in Step 2 from the Sinica corpus (which is a Chinese corpus of two millions words [9]) or other domain specific collections. For example, the following Chinese sentence "醫生的職責就是預防與治療疾病 (A doctor's job is to prevent a disease and to cure the patient)" is a candidate sentence including the Chinese word "疾病 (disease)".

Step 4. Find out all possible verb-senses from the sentences selected in Step 3 to form all possible verb-senses for the selected noun-sense. Count the frequency for each verb-sense.

Step 5. Sort all possible different verb-senses by their corresponding frequencies from large to small. (See Fig. 2) Determine a cut-off frequency in the list. Among all verb-senses above the cut-off frequency, manually pick the permissible ones for the selected noun-sense. Meanwhile, determine their sense subclass symbols (i.e. "&", "%" and "=").

Step 6. Add these permissible NVEF subclass sense-pairs to the KR-tree.

Note that among the above steps, only step 5 requires human intervention. This step is quite laborious, but through learning, human involvement can be greatly reduced. Fig. 2 shows the top 5 possible verb-senses picked by the above algorithm for the noun-sense "disease|疾病" collected from 302 sentences in the Sinica corpus. In Fig. 2, the permissible verb-senses for the noun-sense "disease|疾病" are "cure|醫治" with a frequency of 24, "CauseAffect|傳染,medical|醫" with 23, "ResultIn|導致" with 19 and "obstruct|阻止" with 14. It is observed that, if the number of sentences collected in Step 3 is greater than 300, the top 5 verb-senses will almost always form NVEF sense-pairs with the selected noun-sense.

Figure 2. Top 5 possible verb-senses for creating permissible NVEF sense subclasses for the noun-sense "disease|疾病"

2.4 A Primitive NVEF Sense-Pair Identifier

Based on the KR-tree, we shall develop a primitive NVEF sense-pair identifier as follows. For a given sentence, the algorithm will first identify all NVEF sense-pairs in the KR-tree that have corresponding NVEF word-pairs in the sentence. It will then arrange these NVEF sense-pairs and NVEF word-pairs into a tree, called a *sentence-NVEF tree*, as shown in Fig. 3.

```
(A single-NVEF sentence)
+這輛車行駛順暢 (This car moves well)
+--+N1
+--+--+LandVehicle|車
+--+--+--+車 (car)
+--+V1
+--+--+VehicleGo|駛
+--+--+--+行駛 (move) (the longest syllablic word)
+--+--+--+駛 (move)

(A multi-NVEF sentence)
+趕馬入畜欄 (Drive the horses into the corral)
+--+N1
+--+--+livestock|牲畜
+--+--+--+馬 (horse)
+--+V1
+--+--+expel|驅趕
+--+--+--+趕 (drive)
+--+N2
+--+--+facilities|設施,space|空間,@foster|飼養,#livestock|牲畜
+--+--+--+畜欄 (corral)
+--+V2
+--+--+GoInto|進入
+--+--+--+入 (into)
```

Figure 3. Two sentence-NVEF trees for the input Chinese sentences (a) “這輛車行駛順暢” (a single-NVEF sentence) and (b) “趕馬入畜欄” (a multi-NVEF sentence) respectively

A more formal description of the primitive NVEF sense-pair identifier is given below.

Step 1. Input a sentence.

Step 2. Generate all possible NV word-pairs of the input sentence.

Step 3. Check each NV word-pair to see if their corresponding senses can be matched to a NVEF subclass sense-pair in the KR-tree. If it matches, then use their corresponding noun-sense and verb-sense, respectively for this sentence.

Step 4. Arrange all permissible NVEF sense-pairs and their corresponding NVEF word-pairs in a sentence-NVEF tree.

A system overview of the primitive NVEF identifier is given in Fig. 4.

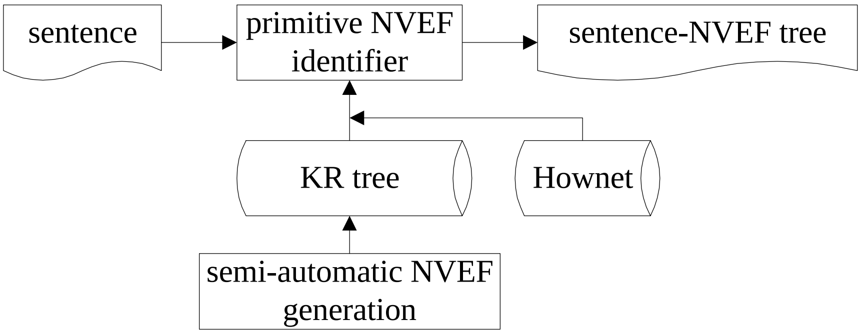

Figure 4. System overview of the primitive NVEF identifier

2.5 A NVEF Sense-Pair Identifier

In Fig. 3, the correct segmented results of the two Chinese sentences are "這/輛/車/行駛/順暢" and "趕/馬/入/畜欄". The upper part of Fig. 3 is a sentence-NVEF tree with a single NVEF sense-pair, "LandVehicle|車"- "VehicleGo|駛", which has two corresponding NV word-pairs, i.e. "車-行駛" and "車-駛". If we further apply the "*longest syllabic NVEF-word-pair first*" strategy (LS-NVWF), the incorrect NVEF word-pair "車-駛" will be successfully dropped. Note that the "longest syllabic word first strategy" is an effective technique for Chinese word segmentation [10]. The lower part of Fig. 3 is a sentence-NVEF tree with two NVEF sense-pairs including "expel|驅趕"-"livestock|牲畜" (NV word-pair is "馬-趕") and "facilities|設施,space|空間,@foster|飼養,#livestock|牲畜"-"GoInto|進入" (NV word-pair is "畜欄-入").

Another useful technique is to exclude certain nouns or verbs from the sentence-NVEF tree. A word with very low frequency as a noun or a verb is treated as a *word of exclusion* for the NVEF sense-pair identifier. Take the Chinese word "的 (of/target)" as an example. Its frequency as a noun or a verb is only 0.004% (computed according to the Sinica corpus). Thus, "的" becomes a word of exclusion. In our experiment, the *exclusion word list* (*EWL*) consists of those words whose frequencies as nouns or verbs are no greater than 5%. When an NVEF word-pair includes at least one exclusion word, its corresponding NVEF sense-pair is excluded from the sentence-NVEF tree. This process is called *EWL checking*. Appendix B lists all exclusion words used in this experiment.

Thus, our final NVEF sense-pair identifier can be described as follows.

Step 1. Input a sentence.

Step 2. Generate all possible NV word-pairs of the input sentence. Exclude certain word-pairs based on EWL checking.

Step 3. Check each NV word-pair to see if their corresponding NV sense-pairs can be matched to a NVEF subclass sense-pair in the KR-tree. For each

NV sense-pair that matches a NVEF subclass sense-pair in the KR-tree, use them as the permissible NVEF sense-pairs, respectively for this sentence. Resolve conflicts using the LS-NVWF strategy.

Step 4. Arrange all permissible NVEF sense-pairs and their corresponding NVEF word-pairs in a sentence-NVEF tree.

A system overview of the NVEF identifier is given in Fig. 5.

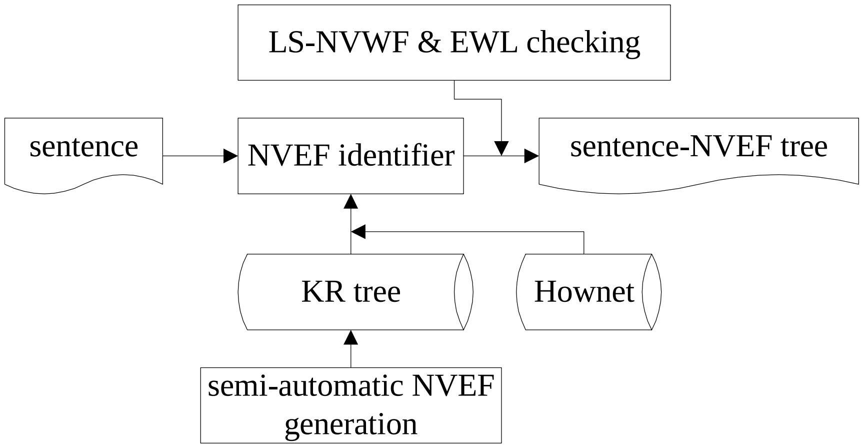

Figure 5. A system overview of the NVEF identifier

To evaluate the WSD performance of the NVEF sense-pair identifier, we consider a WSD experiment in the next section.

3. The WSD experiment

Within a sentence, the number of available NVEF sense-pairs is finite. Consider the Chinese sentence “這輛車行駛順暢 (This car moves well)”. Table 3 gives eight possible pairs of NVEF senses found in this sentence, but there is only one permissible NVEF sense-pair, “LandVehicle|車”-“VehicleGo|駛”.

To evaluate the performance of WSD by using the NVEF sense-pair identifier with the KR-tee, we define the NVEF sense accuracy for a set of test sentences to be

NVEF sense accuracy = # of successful sentences / # of test sentences, (1)

Where a sentence is *successful* if all NVEF sense-pairs and their corresponding NVEF word-pairs obtained from NVEF sense-pair identifier are correct for this sentence. With the KR-tree, the WSD performance for the test sentences can be evaluated by computing the NVEF sense accuracy. This equation is designed from the viewpoint of natural language understanding. Since NVEF sense-pairs often represent a key feature in the meaning of a sentence, any incorrect NVEF sense-pair identification could result in a misunderstanding of this sentence.

Table 3. Eight possible pairs of NVEF senses found in the Chinese sentence “這輛車行駛順暢 (This car moves well)”

C.Word / Noun-sense	C.Word / Verb-sense	P.NVEF [a]
這 / time|時間	行駛 / Go|駛	No
這 / time|時間	車 / cut|切削	No
這 / time|時間	車 / irrigate|澆灌	No
車 / LandVehicle|車	行駛 / VehicleGo|駛	Yes
車 / character|文字,surname|姓,human|人,ProperName|專	行駛 / VehicleGo|駛	No
車 / machine|機器	行駛 / VehicleGo|駛	No
車 / part|部件,%tool|用具,#recreation|娛樂	行駛 / VehicleGo|駛	No
車 / LandVehicle|車	駛 / VehicleGo|駛	Yes

[a] P.NVEF represents permissible NVEF sense-pair

3.1 WSD Evaluation

The framework of WSD evaluation for the NVEF sense-pair identifier is as follows.

1. Select a set of Chinese test sentences from the Sinica Corpus [9] randomly.
2. Use the tool of example-based possible NVEF generation to search and create all permissible NVEF subclass sense-pairs found in these test sentences in the KR-tree.
3. Apply the NVEF sense-pair identifier to these test sentences and obtain their corresponding sentence-NVEF trees
4. Compute the NVEF sense accuracy for the test sentences by Equation 1.

For this study, we have analyzed 7.7% (=764/9,893) of all noun-senses in Hownet and created 4,028 NVEF subclass sense-pairs in the KR-tree. The minimum, maximum and mean number of characters per sentence (of the 445 Chinese test sentences) are 4, 24 and 11.5, respectively. And, the numbers of single-NVEF sentences and multi-NVEF sentences of the test sentences are 96 and 349, respectively.

We conduct the experiment in a progressive manner. The NVEF sense accuracy of the test sentences determined by the NVEF sense-pair identifier using only the knowledge of KR-tree is 74.8% (see Table 4). When the strategy of the longest syllabic NVEF-word-pair first (LS-NVWF) is adopted together with the NVEF sense-pair identifier, the NVEF sense accuracy becomes 87.6%. When the exclusion word list is adopted together with the NVEF sense-pair identifier, the NVEF sense accuracy becomes 89.2%. When both the techniques of LS-NVWF and EWL checking are adopted with the NVEF sense-pair identifier (see Table 4), the NVEF sense accuracy is improved to 93.7%. Along with the NVEF sense-pair identifier, the word-segmentation accuracy (for those ambiguous NVEF word-pairs) for these sentences is 99.6%(443/445). This result also supports the aforementioned claim that the NVEF word-segmentation accuracy is better than the NVEF sense accuracy. Appendix C gives two successful and one unsuccessful sentence-NVEF trees in this experiment.

Table 4. Results of the WSD experiment for 445 Chinese un-segmented test sentences

# of NVEF	NVEF sense accuracy	Using LS-NVWF [a]	Using EWL[b]	Using Both[c]
4,028	74.8%(333/445)	87.6%(390/445)	89.2%(397/445)	93.7%(417/445)

[a] "Using LS-NVWF" represents NVEF sense accuracy by using LS-NVWF with the NVEF sense-pair identifier
[b] "Using EWL" represents NVEF sense accuracy by using EWL checking with the NVEF sense-pair identifier
[c] "Using Both" represents NVEF sense accuracy by using both LS-NVWF and EWL checking with the NVEF sense-pair identifier

3.2 An Analysis of the Unsuccessful Cases

Although the NVEF sense accuracy can be up to 93.7% by adopting both the techniques of LS-NVWF and EWL checking with the NVEF sense-pair identifier, there is still a room for improvement. Below, we have classified the reasons behind those unsuccessful cases into four major types:

(1) ***Lack of a bottom-up tagger:*** There are many specific linguistic units such as names, addresses, determinative-measure compounds, etc. in a sentence, which need to be recognized in order to supplement the NVEF sense-pair identifier (which works in a top-down fashion). In this study, 6 unsuccessful sentences are caused by this reason. Although the techniques of LS-NVWF and EWL checking have inadvertently resolved these cases, it is still a potential problem.

(2) ***Lack of Non-NVEF knowledge:*** Consider the Chinese sentence, "太太要掌握先生的荷包 (A wife wants to take her husband's wallet into her hands)". There are three different noun-senses of the Chinese word, "先生(teacher/doctor/husband)", which can form NVEF sense-pair with the verb-sense "掌握(take…into one's hands)". To get the correct noun-sense "先生(husband)" for this sentence, we need the knowledge of noun-noun (NN) sense-pair, such as "太太(wife)"-to-"先生(husband)" or other contextual information. This knowledge is not available from the KR-tree and needs to be collected separately. In this study, there are 15 unsuccessful sentences due to this reason, which cannot be resolved by the techniques of LS-NVWF or EWL checking.

(3) ***Inadequate word segmentation:*** Consider the Chinese sentence, "他以滿分得到冠軍(He received the championship with a full mark)". There are two possible verbs with the same verb-sense "分得(obtain)" and "得到(obtain)" that can form

NVEF sense-pairs with the noun-sense “冠軍(champ)”. In this case, we have two conflicting NVEF sense-pairs and need a better segmentation algorithm to determine that the correct verb is “得到(obtain)” for this sentence (the correct segmented results of this sentence is “他/以/滿分/得到/冠軍”). In this study, there are 3 unsuccessful sentences due to this reason, which cannot be resolved by the techniques of LS-NVWF or EWL checking.

(4) ***Lack of a multi-NVEF analyzer:*** Consider the Chinese sentence, “搭飛機離開台北(Take airplane to leave Taipei)”. The NVEF sense-pair identifier detects that there are three NVEF sense-pairs: N_1-V_1: [N_1=飛機(airplane),V_1=搭(take)], N_2-V_2: [N_2=台北(Taipei),V_2=離開(leave)], and N_3-V_3: [N_3=飛機(airplane),V_3=離開(leave)] in the sentence. In this case, N_1-V_1 and N_2-V_2 can be used to construct a permissible bi-NVEF sequence V_1-$N_1$$V_2$-$N_2$, which will be competing with the NVEF sense pair N_3-V_3. Currently, such cases are not analyzed since our system does not yet have the knowledge of permissible multi-NVEF sense-pairs. In this study, there are 5 incorrect sentences due to this reason, which cannot be resolved by the techniques of LS-NVWF or EWL checking.

In case these four problems can be resolved, the NVEF sense accuracy can be improved to (417+15+3+5)/(445) = 98.9%.

Based on this experiment, we find that our NVEF sense-pair identifier has the potential to provide the following information for a given sentence: (1) main verbs, (2) nouns, (3) NVEF word-pairs, (4) NVEF sense-pairs, (5) NVEF phrase-boundaries, and (6) initial relationship of multi-NVEF sense/word-pairs. A correct NVEF sense-pair will naturally include the correct NVEF word-pair for word segmentation. However, the converse is not true. Namely, a correct NVEF word-pair cannot be used to assure that the corresponding NVEF sense-pair is permissible. Thus, the NVEF word-segmentation accuracy is normally better than the NVEF sense accuracy.

4. Conclusion and Future Direction

In this paper, we describe a NVFE sense-pair identifier with an attempt to disambiguate word sense in Chinese sentences. A WSD experiment was conducted by using the NVEF sense-pair identifier with the KR-tree. The knowledge in the KR-tree is created with the help of a semi-automatic NVEF generation tool.

Based on current techniques, our experiment shows that the NVEF sense accuracy is 93.7% and the NVEF word-segmentation accuracy is 99.6%. We have indicated, in Section 3, several directions that can further improve the performance of our system, some of which are currently being studied.

Our experiment indicates that the NVEF sense-pair knowledge is effective for NVEF word-sense disambiguation in Chinese sentences. It also supports the claim in

[5] that, people usually disambiguate word sense with only a few words (frequently only one word) in the context. We are particularly pleased to note that the NVEF knowledge can achieve a high accuracy for NVEF word-segmentation since a correct word-segmentation is a primary key for a successful Chinese NLP [11].

Although we have a semi-automatic NVEF generation tool, it is still laborious to create our current NVEF knowledge, which constitutes only 7.7% of the entire NVEF knowledge. Hence, a systematic method for fully automatic NVEF knowledge generation is highly desirable. Furthermore, we will try to develop a combined top-down and bottom-up NVEF sense-pair identifier that can address the issues involving the four unsuccessful cases in Section 3.

We plan to create a full fletch KR-tree so that we can investigate the robustness of the sense-based approach for monolingual and bilingual (e.g. English-Chinese) WSD. The study of NVEF will also be extended to noun-noun pairs, noun-adjective pairs as well as verb-adverb pairs. Another related research topic is to apply the NVEF sense-pair identifier to other fields of NLP, in particular, document classification, information retrieval, question answering and speech understanding.

5. Acknowledgement

We are grateful to the following colleagues in the Intelligent Agent Systems Lab., Li-Yeng Chiu, Mark Shia, Gladys Hsieh, Masia Yu and Yi-Fan Chang, who help us create all the necessary NVEF knowledge for this study. We would also like to thank Prof. Zhen-Dong Dong for providing us the Hownet dictionary.

Reference

1. Weaver, W. *Translation*. Mimeographed, 1949, pp. 12.
2. Small, S., Cottrell, G., Tannenhaus, M., Eds, *Lexical Ambiguity Resolution*. Morgan Kaufmann, Palo Alto, Calif., 1988.
3. Krovetz, R. and Croft, W. B. Lexical Ambiguity and Information Retrieval, *ACM Transactions on Information Systems*. 10 (2), 1992, pp.115-141.
4. Resnik, P. and Yarowsky, D. Distinguishing Systems and Distinguishing Senses: New Evaluation Methods for Word Sense Disambiguation. *Natural Language Engineering.* 5 (3), 2000, pp.113-133.
5. Choueka, Y. and Lusignan, S. A Connectionist Scheme for Modeling Word Sense Disambiguation. *Cognition and Brain Theory*, 6 (1), 1983, pp.89-120.
6. Carey, S., The origin and evolution of everyday concepts. In R. N. Giere, ed. *Cognitive Models of Science*. Minneapolis: University of Minnesota Press, 1992.
7. Dong, Z. and Dong, Q., Hownet, http://www.keenage.com/
8. Voorhees, E. M. Using WordNet to Disambiguate Word Senses for Text

Retrieval. *ACM-SIGIR*, 1993.

9. CKIP. *Technical Report no. 95-02, the content and illustration of Sinica corpus of Academia Sinica*. Institute of Information Science, Academia Sinica, 1995. http://godel.iis.sinica.edu.tw/CKIP/r_content.html
10. Chen, C.-G., Keh-Jiann Chen and Lin-Shan Lee. A model for Lexical Analysis and Parsing of Chinese Sentences. *Proceedings of 1986 International Conference on Chinese Computing.* Singapore, 1986, pp.33-40.
11. Slocum, J. and Justus, C. F. Transportability to Other Language: The Natural Language Processing Project in the AI Program at MCC. ACM Transactions on Office Information Systems, 3(2) 1985, pp.204-230.

Appendix A. A Sample Table of Main Features of Nouns its corresponding Noun-Sense Classes

An example Main Feature	Noun-sense Class
bacteria|微生物	微生物
AnimalHuman|動物	動物類
human|人	人物類
plant|植物	植物類
artifact|人工物	人工物
natural|天然物	天然物
fact|事情	事件類
mental|精神	精神類
phenomena|現象	現象類
shape|物形	物形類
InstitutePlace|場所	地點類
location|位置	位置類
attribute|屬性	抽象類
quantity|數量	數量類

Appendix B. Exclusion Word List

I. Monosyllabic exclusion words
/之/的/不/與/兩/再/以/了/較/就/次/得/於/已/把/都/太/一/某/最/
/內/均/原/由/被/全/初/及/將/該/總/塊/項/和/二/從/三/凡/尚/前/
/十/極/番/元/件/甚/因/甲/向/才/四/本/若/先/便/五/粒/常/卅/後/
/左/曾/竟/廿/八/支/六/著/首/剛/應/篇/能/七/終/依/位/暫/共/須/
/中/九/時/可/俱/整/謹/宜/邊/往/批/夥/在/唔/年/諸/略/東/特/磅/

II. Polysyllabic exclusion words
/所以/不能/不會/是否/之間/終於/不必/唯一/西方/恐怕/連續/
/必須/不妨/大家/不得/一旦/初步/據說/看來/全面/臨床/無數/
/依法/國立/過度/突然/通常/一同/單一/大力/純粹/大都/當然/
/種種/大概/國有/順便/總是/不再/默默/無不/那麼/黑白/個人/
/四處/自行/恰好/終究/最佳/一心/十分/甚為/私立/一起/可以/
/多元/所有/依然/現成/正好/針對/一般/難怪/等到/到底/應該/
/貿然/獨家/原先/根據/微微/不勝/國產/整整/衷心/好些/安然/
/慈善/為什麼/一下子/一塊兒/非正式/

Appendix C. Three sentence-NVEF trees in this study

I. Successful sentence-NVEF tree

+林震南舉起手中煙袋 (Lin Cheng-Nan uptakes the pipe on his hand)
+--+N1
+--+--+tool|用具,*addict|嗜好
+--+--+--+煙袋 (pipe)
+--+V1
+--+--+lift|提昇
+--+--+--+舉起 (uptake)

II. Successful sentence-NVEF tree

+我所下的部分結論 (Parts of conclusion I have given)
+--+N1
+--+--+thought|念頭,$decide|決定
+--+--+--+結論 (conclusion)
+--+V1
+--+--+announce|發表/V
+--+--+--+下 (give)

III. Unsuccessful sentence-NVEF tree including one incorrect word sense "樹(put)"

+樹上掛著黃絲帶 (Yellow silk ribbons are hanged on the tree)
+--+N1
+--+--+tool|用具,linear|線,*fasten|拴連,*decorate|裝飾
+--+--+--+絲帶 (silk ribbon)
+--+V1
+--+--+put|放置
+--+--+--+樹 (put)
+--+N2
+--+--+tool|用具,linear|線,*fasten|拴連,*decorate|裝飾
+--+--+--+絲帶 (silk ribbon)
+--+V2
+--+--+ hang|懸掛
+--+--+--+掛 (hang)