Auto-Discovery of NVEF Word-Pairs in Chinese

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

Abstract

A meaningful noun-verb word-pair in a sentence is called a noun-verb event-frame (NVFE). Previously, we have developed an NVEF word-pair identifier to demonstrate that NVEF knowledge can be used effectively to resolve the Chinese word-sense disambiguation (WSD) problem (with 93.7% accuracy) and the Chinese syllable-to-word (STW) conversion problem (with 99.66% accuracy) on the NVEF related portion.

In this paper, we propose a method for automatically acquiring a large scale NVEF knowledge without human intervention. The automatic discovery of NVEF knowledge includes four major processes: (1) segmentation check; (2) Initial Part-of-speech (POS) sequence generation; (3) NV knowledge generation and (4) automatic NVEF knowledge confirmation.

Our experimental results show that the precision of the automatically acquired NVEF knowledge reaches 98.52% for the test sentences. In fact, it has automatically discovered more than three hundred thousand NVEF word-pairs from the 2001 *United Daily News* (2001 *UDN*) corpus. The acquired NVEF knowledge covers 48% NV-sentences in *Academia Sinica Balanced Corpus* (***ASBC***), where an NV-sentence is one including at least a noun and a verb.

In the future, we will expand the size of NVEF knowledge to cover more than 75% of NV-sentences in ***ASBC***. We will also apply the acquired NVEF knowledge to support other NLP researches, in particular, shallow parsing, syllable/speech understanding and text indexing.

Keywords: noun-verb event frame (NVEF), machine learning, Hownet, WSD, STW

1. Introduction

The most challenging problem in NLP is to program computers to understand natural languages. For a human being, efficient syllable-to-word (STW) conversion and word sense disambiguation (WSD) arise naturally while a sentence is understood. Therefore, in designing a natural language understanding (NLD) system, two basic problems are to derive methods and knowledge for effectively performing the tasks of STW and WSD.

For most languages, a sentence is a grammatical organization of words expressing a complete thought [Chu 1982, Fromkin *et al*. 1998]. Since a word is usually encoded with ploy-senses, to understand language, efficient word sense disambiguation (WSD) becomes a critical problem for any NLD system. According to a study in cognitive science [Choueka *et al*. 1983], people often disambiguate word sense using only a few other words in a given context (frequently only one additional word). Thus, the relationships between one word and others can be effectively used to resolve ambiguity. Furthermore, from [Small *et al*. 1988, Krovetz *et al*. 1992, Resnik *et al*. 2000], most ambiguities occur with nouns and verbs, and the object-event (i.e. noun-verb) distinction is a major ontological division for humans [Carey 1992]. Tsai *et al*. (2002) have shown that the knowledge of noun-verb event frame (NVEF) sense/word-pairs can be used effectively to achieve a WSD accuracy of 93.7% for the NVEF related portion in Chinese, which supports the above claim of [Choueka *et al*. 1983].

The most common relationships between nouns and verbs are subject-predicate (SP) and verb-object (VO) [胡裕樹 *et al*. 1995, Fromkin *et al*. 1998]. In Chinese, such NV relationships could be found in various language units: compounds, phrases or sentences [Li *et al*. 1997]. As our observation, the major NV relationships in compounds/phrases are SP, VO, MH (modifier-head) and VC (verb-complement) constructions; the major NV relationships in sentences are SP and VO constructions. Consider the Chinese sentence: **這輛車行駛順暢**(This car moves well). There are two possible NV word-pairs, "車-行駛(car, move)" and "車行-駛(auto shop, move)." It is clear that the permissible (or meaningful) NV word-pair is "車-行駛(car, move)" and it is a SP construction. We call such a permissible NV word-pair a noun-verb *event frame* (*NVEF*) word-pair. And, the collection of the NV word-pair **車-行駛** and its sense-pair **Land-Vehicle|車-VehicleGo|駛** is called a permissible NVEF knowledge.

The most popular input method for Chinese is syllable-based. Since the average number of characters sharing the same syllable is 17, efficient STW conversion becomes an indispensable tool. Tsai *et al*. (2002) have shown that the NVEF knowledge can be used to achieve a STW accuracy rate of 99.66% for converting NVEF related words. Since the creation of NVEF knowledge bears no particular application in mind, and still it can be used to effectively resolve the WSD and STW problems, the NVEF knowledge is potentially application independent for NLP. We shall further investigate the effectiveness of NVEF knowledge in other NLP applications,

such as syllable/speech understanding and full/shallow parsing.

We have reported a semi-automatic generation of NVEF knowledge in [Tsai *et al*. 2002]. This method uses the N-V frequencies in sentences groups to generate NVEF candidates to be filtered by human editors. However, it is quite laborious to create a large scale NVEF knowledge. In this paper, we propose a new method to discover NVEF knowledge automatically from running texts, and construct a large scale NVEF knowledge efficiently.

This paper is arranged as follows. In Section 2, we present the details of auto-discovery of NVEF knowledge. Experimental results and analyses are described in Section 3. Conclusion and directions for future researches will be discussed in Section 4.

2. Development of Auto-Discovery of NVEF Knowledge

To develop an auto-discovery system for NVEF knowledge (AUTO-NVEF), we use Hownet 1.0 [Dong] as a system dictionary. This system dictionary provides knowledge of the Chinese word (58,541 words), parts-of-speech (POS) and word senses, in which there are 33,264 nouns, 16,723 verbs and 16,469 senses (including 10,011 noun-senses and 4,462 verb-senses).

2.1 Definition of the NVEF Knowledge

The sense of a word is defined as 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 the 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

As we mentioned, a permissible (or meaningful) NV word-pair is a noun-verb event-frame

word-pair (*NVEF word-pair*), such as **車-行駛**(Che/car/turn, move). From Table 2, the only permissible NVEF sense-pair for **車-行駛**(car, move) is **LandVehicle|車-VehicleGo|駛**. Such an NVEF sense-pair and its corresponding NVEF word-pairs is called NVEF knowledge. Here, the combination of the NVEF sense-pair **LandVehicle|車-VehicleGo|駛** and the NVEF word-pair **車-行駛** constructs a collection of NVEF knowledge.

To effectively represent the NVEF knowledge, we have proposed an NVEF knowledge representation tree (NVEF KR-tree) to store and display the collected NVEF knowledge. The details of the NVEF KR-tree are described below [Tsai *et al*. 2002].

2.2 Knowledge Representation Tree of NVEF Sense-Pairs and Word-Pairs

A knowledge representation tree (KR-tree) of NVEF sense-pairs is shown in Fig.1.

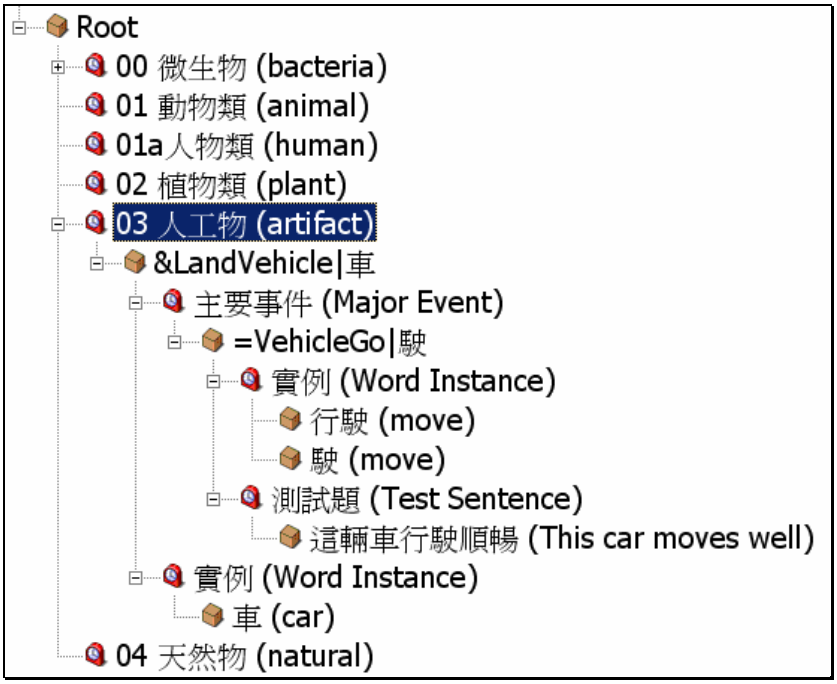

Figure 1. An illustration of the KR-tree using "人工物(artifact)" as an example noun-sense subclass. (The English words in parentheses are provided for explanatory purposes only.)

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 the parent and children concept nodes. We omit the function node "subclass" so that if a concept node B is the child of another concept node A, then B is a subclass of A. We can classify the noun-sense class (名詞詞義分類) into 15 subclasses according to their main features. These 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 provides a sample table of the 15 main features of nouns in each noun-sense subclass.

The three function nodes used in the KR-tree are shown in Figure 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 an NVEF subclass sense-pair, such as "&LandVehicle|車" and "=VehcileGo|駛" in Figure 1. To describe various relationships between noun-sense and verb-sense subclasses, we design three subclass sense-symbols, in which "=" means "*exact*," "&" means "*like*," and "%" means "*inclusive*." An example using these symbols is provided below.

Provided that there are three senses S_1, S_2, and S_3 as well as their corresponding words W_1, W_2, and W_3. Let

S_1 = LandVehicle|車,*transport|運送,#human|人,#die|死 W_1="靈車(hearse)"

S_2 = LandVehicle|車,*transport|運送,#human|人 W_2="客車(bus)"

S_3 = LandVehicle|車,police|警 W_3="警車(police car)"

Then, we have that sense/word S_3/W_3 is in the "=LandVehicle|車,police|警" *exact*-subclass; senses/words S_1/W_1 and S_2/W_2 are in the "&LandVehicle|車,*transport|運送" *like*-subclass; and senses/words S_1/W_1, S_2/W_2, and S_3/W_3 are in the "%LandVehicle|車" *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 sense-pair 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 Auto-Discovery of NVEF Knowledge

The task of AUTO-NVEF is to automatically find out meaningful NVEF sense/word-pairs (NVEF knowledge) from Chinese sentences. Figure 1 is the flow chart of AUTO-NVEF. There are four major processes in AUTO-NVEF. The details of these major processes are described as follows (see Figure 2 and Table 2).

Process 1. Segmentation check: In this stage, the Chinese sentence will be segmented by two strategies: *right-to-left longest word first* (RL-LWF), and *left-to-right longest word first* (LR-LWF). If both RL-LWF and LR-LWF segmentations are equal (in short form, RL-LWF=LR-LWF) and the word number of the segmentation is greater than one, this segmen-

tation result will be sent to ***process 2***; otherwise, a *NULL* segmentation will be sent. Table 3 is a comparison of word-segmentation accuracies for RL-LWF, LR-LWF and RL-LWF=LR-LWF strategies with CKIP lexicon [CKIP 1995]. The word-segmentation accuracy is the ratio of fully correct segmented sentences to all sentences of *Academia Sinica Balancing Corpus* (***ASBC***) [CKIP 1995]. A fully correct segmented sentence means the segmented result exactly matches its corresponding segmentation ***ASBC***. Table 3 shows that the technique of RL-LWF=LR-LWF achieves the best word-segmentation accuracy.

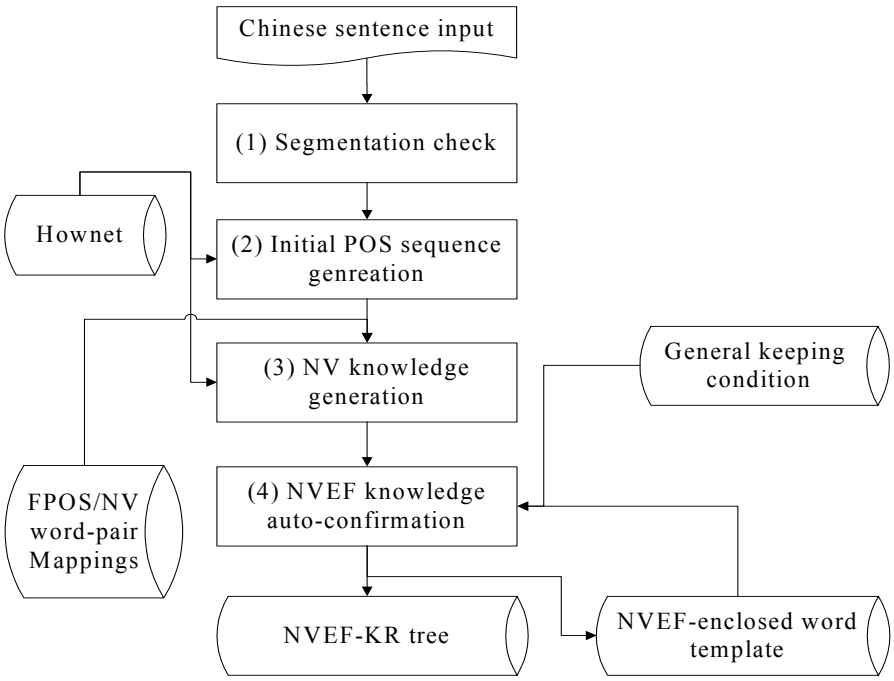

Figure 2. The flow chart of AUTO-NVEF

Table 2. An illustration of AUTO-NVEF for the Chinese sentence "音樂會現場湧入許多觀眾 (There are many audiences entering the locale of concert)." (The English words in parentheses are included for explanatory purpose only.)

Process	Output
(1)	音樂會(concert)/現場(locale)/湧入(enter)/許多(many)/觀眾(audience)
(2)	$N_1N_2V_3ADJ_4N_5$, where N_1 =[音樂會]; N_2 =[現場]; V_3=[湧入]; ADJ_4=[許多]; N_5=[觀眾]
(3)	NV_1 = "現場/place|地方,#fact|事情/N" - "湧入(yong3 ru4)/GoInto|進入/V" NV_2 = "觀眾/human|人,*look|看,#entertainment|藝,#sport|體育,*recreation|娛樂/N" - "湧入(yong3 ru4)/GoInto|進入/V"
(4)	NV_1 is NVEF knowledge by keeping-condition; learned NVEF template is [音樂會 NV 許多] NV_2 is NVEF knowledge by keeping-condition; learned NVEF template is [現場 V 許多 N]

Table 3. A comparison of word-segmentation accuracies for RL-LWF, LR-LWF and RL-LWF = LR-LWF strategies (the test sentences are ***ASBC*** and the dictionary is CKIP lexicon)

	RL-LWF	LR-LWF	RL-LWF = LF-LWF
Accuracy	82.5%	81.7%	86.86%
Recall	100%	100%	89.33%

Process 2. Initial POS sequence generation: If the output of ***process 1*** is not a *NULL* segmentation, this process will be triggered. This stage is comprised of the following steps.

1) For the segmentation result $w_1/w_2/\ldots/w_{n-1}/w_n$ from ***process 1***, our algorithm compute the POS of w_i, where i = 2 to n, as follows. It first computes the following two sets: a) the *following POS/frequency set* of w_{i-1} by ***ASBC*** tagging corpus and b) the *Hownet POS set* of w_i. Then, it computes the POS intersection of the two sets. Finally, it selects the POS with the largest frequency in the POS intersection to be the POS of w_i. If there are more than one POS with the largest frequency, the POS of w_i will be set to *NULL* POS.
2) Similarly, the POS of w_1 will be determined by the POS with the largest frequency in the POS intersection of the *preceding POS/frequency set* of w_2 and the *Hownet POS set* of w_1.
3) By combining the determined POSs of w_i, where *i* =1 to n, the ***initial POS sequence*** (***IPOS***) will be generated. Take the Chinese segmentation 生/了 as an example. The following POS/frequency set of the Chinese word 生(bear) is {N/103, PREP/42, STRU/36, V/35, ADV/16, CONJ/10, ECHO/9, ADJ/1}. The Hownet POS set of the Chinese word 了 is {V, STRU}. According to these sets, we have POS intersection {STRU/36, V/35}. Since the POS with the largest frequency in this intersection is **STRU**, the POS of 了 will be set to **STRU**. Similarly, according to the intersection {V/16124, N/1321, ADJ/4} of the preceding POS/frequency set {V/16124, N/1321, PREP/1232, ECHO/121, ADV/58, STRU/26, CONJ/4, ADJ/4} of 了 and the Hownet POS set {V, N, ADJ} of 生, the POS of 生 will be set to **V**. Table 4 is a mapping list of CKIP POS tag and Hownet POS tag.

Table 4. A mapping list of CKIP POS tag and Hownet POS tag

	Noun	Verb	Adjective	Adverb	Preposition	Conjunction	Expletive	Structural Particle
CKIP	N	V	A	D	P	C	T	De
Hownet	N	V	ADJ	ADV	PP	CONJ	ECHO	STRU

Process 3. NV knowledge generation: If the output of ***process 2*** does not include any *NULL* POS, this process will be triggered. The steps of this process are given as follows.

1) Compute the ***final POS sequence*** (***FPOS***). For the portion of contiguous noun sequence (such as N_1N_2) of the ***IPOS***, the last noun (such as N_2) will be kept and the other nouns will

be dropped from the ***IPOS***. This is because the last noun of a contiguous noun sequence (such as 航空/公司) in Chinese is usually the head of such a sequence. This step translates an ***IPOS*** into a ***FPOS***. Take the Chinese sentence 音樂會(N_1)現場(N_2)湧入(V_3)許多(ADJ_4)觀眾(N_5) as an example. Its ***IPOS*** ($N_1N_2V_3ADJ_4N_5$) will be translated into ***FPOS*** ($N_1V_2ADJ_3N_4$).

2) According to the ***FPOS***, the NV word-pairs will be generated. In this case, since the auto-generated NV word-pairs for the ***FPOS*** $N_1V_2ADJ_3N_4$ are N_1V_2 and N_4V_2, the NV word-pairs **現場(N)湧入(V)** and **湧入(V)觀眾(N)** will be generated. Appendix. B lists three sample mappings of the ***FPOSs*** and their corresponding NV word-pairs. In this study, we create about one hundred mappings of ***FPOSs*** and their corresponding NV word-pairs.
3) According to Hownet, it computes all NV sense-pairs for the generated NV word-pairs. For the above case, we have two collections of NV knowledge (see Table 2):

NV_1 = "現場(locale)/place|地方,#fact|事情/N" – "湧入(enter)/GoInto|進入/V", and

NV_2 = "觀眾(audience)/human|人,*look|看,#entertainment|藝,#sport|體育,*recreation|娛樂/N" – "湧入(enter)/GoInto|進入/V".

Process 4. NVEF knowledge auto-confirmation: In this stage, it automatically confirms whether the generated NV knowledge is NVEF knowledge. The two auto-confirmation procedures are given as follows.

(a) **General keeping (GK) condition check**: Each GK condition is constructed by a noun-sense class defined in [Tsai *et al*. 2002] (see Appendix A) and a verb main DEF in Hownet 1.0 [Dong]. For example, the pair of noun-sense class "人物類(human)" and verb main DEF "GoInto|進入" is a GK condition. In [Tsai *et al*. 2002], we created 5,680 GK conditions from the manually confirmed NVEF knowledge. If the noun-sense class and the verb main DEF of the generated NV knowledge fits one of GK conditions, it will be automatically confirmed as a collection of NVEF knowledge and sent to NVEF KR-tree. Appendix. C gives ten GK conditions used in this study.

(b) **NVEF enclosed-word template (NVEF-EW template) check**: If the generated NV knowledge cannot be auto-confirmed as NVEF knowledge in procedure (a), this procedure will be triggered. A NVEF-EW template is composed of all left words and right words of a NVEF word-pair in a Chinese sentence. For example, the NVEF-EW template of the NVEF word-pair "汽車-行駛(car, move)" in the Chinese sentence 這(this)/汽車(car)/似乎(seem)/行駛(move)/順暢(well) is ***這 N 似乎 V 順暢***. In this study, all the NVEF-EW templates are generated from the following resources: i) the collection of manually confirmed NVEF knowledge in [Tsai *et al*. 2002], ii) the automatically confirmed NVEF knowledge and iii) the NVEF-EW templates provided by human editor. In this procedure, if the NVEF-EW template of the generated NV word-pair for the Chinese sentence input matches one of the NVEF-EW templates, it will be automatically confirmed as a col-

lection of NVEF knowledge.

3. Experiments

To evaluate the performance of the proposed auto-discovery of NVEF knowledge, we define the NVEF accuracy and NVEF-identified sentence coverage by Equations (1) and (2):

NVEF accuracy =

# of permissible NVEF knowledge / # of total generated NVEF knowledge. (1)

NVEF-identified sentence coverage =

# of NVEF-identified sentences / # of total NV sentences. (2)

In Equation (1), a permissible NVEF knowledge means the generated NVEF knowledge is manually confirmed as a collection of NVEF knowledge. In Equation (2), if the Chinese sentence contains greater or equal to one NVEF word-pair on our NVEF KR-tree by the NVEF word-pair identifier [Tsai *et al*. 2002], this sentence is called an **NVEF-identified sentence**. If the Chinese sentence contains at least one noun and verb, this sentence is called an **NV sentence**. As our computation, there are about 75% of Chinese sentences in Sinica corpus are NV sentences.

Chinese sentence	高度壓力使有些人[食量]<減少> (High pressure makes some people that their [eating capacity] <decreased>.)		
名詞詞義 (Noun sense)	attribute|屬性,ability|能力,&eat|吃	動詞詞義 (Verb sense)	subtract|削減
名詞 (Noun)	食量 (eating capacity)	動詞 (Verb)	減少 (decrease)

Figure 3. The confirmation UI of NVEF knowledge taking the generated NVEF knowledge for the Chinese sentence 高度壓力使有些人食量減少 (High pressure makes some people that their eating-capacity decreased as an example. (The English words in parentheses, symbols [] used to mark a noun and <> used to mark a verb are there for explanatory purposes only)

3.1 User Interface (UI) for Manually Confirming NVEF Knowledge

An evaluation UI for the generated NVEF knowledge is developed as shown in Figure 3. By this UI, evaluators (native Chinese speakers) can review the generated NVEF knowledge and determine whether it is a permissible NVEF knowledge. Take the Chinese sentence 高度壓力使有些人食量減少(High pressure makes some people that their eating capacity decreased) as an

example. For this case, AUTO-NVEF will generate a collection of NVEF knowledge including the NVEF sense-pair [attribute|屬性,ability|能力,&eat|吃]-[subtract|削減] and the NVEF word-pair [食量(eating capacity)]-[減少(decrease)]. According to the confirmation principles of permissible NVEF knowledge, evaluators will confirm this generated NVEF knowledge as a permissible NVEF knowledge. The confirmation principles of permissible NVEF knowledge are given as follows.

3.2 Confirmation Principles of permissible NVEF Knowledge

An auto-generated NVEF knowledge should be confirmed as a collection of permissible NVEF knowledge if it fits all three principles below.

Principle 1. Do the NV word-pair make correct POS tags for the given Chinese sentence?

Principle 2. Do the NV sense-pair and the NV word-pair make sense?

Principle 3. Do most NV word-pair instances for the NV sense-pair satisfy Principles 1 and 2?

3.3 Experimental Results

To evaluate the acquired NVEF knowledge, we divide the 2001 *United Daily News* (2001 UDN) corpus into two distinct sub-corpora. (The UDN 2001 corpus contains 4,539,624 Chinese sentences that were extracted from the *United Daily News* Web site [On-Line United Daily News] from January 17, 2001 to December 30, 2001.)

(1) **Training corpus.** This is the collection of Chinese sentences extracted from the 2001 *UDN* corpus from January 17, 2001 to September 30, 2001. According to the training corpus, we create thirty thousand manually confirmed NVEF word-pairs, which are used to derive the 5,680 general keeping conditions.

(2) **Testing corpus.** This is the collection of Chinese sentences extracted from the 2001 *UDN* corpus from October 1, 2001 to December 31, 2001.

(3) **Test sentences set**. From the testing corpus, we randomly select three days' sentences (October 27, 2001, November 23, 2001 and December 17, 2001) to be our test sentences set.

All of the acquired NVEF knowledge by AUTO-NVEF on the test sentences are manually confirmed by three evaluators. Table 5 is the experimental results of AUTO-NVEF. From Table 5, it shows that AUTO-NVEF can achieve a NVEF accuracy of 98.52%.

Table 5. Experimental results of AUTO-NVEF

Date of test news	NVEF accuracy	Evaluator
October 27, 2001	99.10% (1,095/1,105)	A
November 23, 2001	97.76% (1,090/1,115)	B
December 17, 2001	98.63% (2,156/2,186)	C
Total Average	98.52% (4,341/4,406)	

When we apply AUTO-NVEF to the entire 2001 UDN corpus, it auto-generates 167,203 NVEF sense-pairs (8.6M) and 317,820 NVEF word-pairs (10.1M) on the NVEF KR-tree. Within this data, 47% is generated through the general keeping conditions check and the other 53% is generated by the NVEF-enclosed word templates check.

Table 6. An illustration of four types of NVEF knowledge and their coverage (The English words in parentheses, symbols [] and <> are there for explanatory purposes only)

NV pair Type	Sentence	Noun / DEF	Verb / DEF	Coverage
N:V	**[工程]<完成>** (The construction is now completed)	工程 (construction) affairs|事務,industrial|工	完成 (complete) fulfil|實現	24.15%
N-V	全部**[工程]**預定年底**<完成>** (All of constructions will be completed by the end of year)	工程 (construction) affairs|事務,industrial|工	完成 (complete) fulfil|實現	43.83%
V:N	**<完成>[工程]** (to complete a construction)	工程 (construction) affairs|事務,industrial|工	完成 (complete) fulfil|實現	19.61%
V-N	建商承諾在年底前**<完成>**鐵路**[工程]** (The building contractor promise to complete railway construction before the end of this year)	工程 (construction) affairs|事務,industrial|工	完成 (complete) fulfil|實現	12.41%

3.3.1 Coverage for the Four Types of NVEF Knowledge

According to the noun and verb positions of NVEF word-pairs in Chinese sentences, the NVEF knowledge can be classified into four types: **N:V**, **N-V**, **V:N**, and **V-N**, where the symbols ":" stands for "next to" and "-" stands for "near by." Table 6 shows examples and the coverage of the four types of NVEF knowledge, in which the ratios (coverage) of the collections of **N:V**, **N-V**, **V:N** and **V-N** are 12.41%, 43.83%, 19.61% and 24.15%, respectively, by applying AUTO-NVEF to 2001 *UDN* corpus. It seems that the percentage of **SP** construction is a little more than that of **VO** construction in the training corpus.

3.3.2 Error Analysis - The Non-Permissible NVEF Knowledge Generated by AUTO-NVEF

One hundred collections of the generated non-permissible NVEF (NP-NVEF) knowledge are analyzed. We classify these into eleven error types as shown in Table 7, which lists the NP-NVEF confirmation principles and the ratios for the eleven error types. The first three types

consist of 52% of the cases that do not satisfy the NVEF confirmation principles 1, 2 and 3 in Section 3.2. The fourth type is rare with 1% of the cases. Types 5 to 7 consists of 11% of the cases and are caused from incorrect Hownet lexicon, such as the incorrect word-sense ***exist|存在*** for the Chinese word ***盈盈*** (an adjective, normally used to describe a beauty's smile). Types 8 to 11 are referred to as the *four NLP errors* (36% of NP-NVEF cases): Type 8 is the problem of different word-senses used in Ancient and Modern Chinese; type 9 is caused by errors in WSD; type 10 is caused by the unknown word problem; and type 11 is caused by incorrect word segmentation.

Table 7. The eleven error types and their confirming principles of non-permissible NVEF knowledge generated by AUTO-NVEF

Type	**Confirming principle of Non-Permissible NVEF Knowledge**	Percentage
1*	NV Word-pair cannot make a reasonable and legitimate POS tagging for the Chinese sentence.	33% (33/100)
2*	NV sense-par (DEF) and the NV word-pair cannot make sense for each other	17% (17/100)
3*	In this NV pair, one of word sense cannot inherit its parent category.	2% (2/100)
4**	The NV pair cannot be the proper combination in the sentence although this pair fits principles (a), (b), and (c).	1% (1/100)
5	Incorrect word POS in Hownet	1% (1/100)
6	Incorrect word sense in Hownet	3% (3/100)
7	No proper definition in Hownet Ex:暫居(temporary residence)，it has two meanings, one is <reside|住下>（緊急暫居服務(Emergent temporary residence service)）and another one is <situated|處,Timeshort|暫>（SARS 帶來暫時性的經濟震盪(SARS will produce only a temporary economic shock))．	7% (7/100)
8	Lack of different meaning usage for Old Chinese and Modern Chinese	3% (3/100)
9	Failure of word sense disambiguation (1) General sense Polysemous word (2) Domain sense Person name, Appellation, Organization named as common word Ex: 公牛隊(**Chicago Bulls)** ⇨公牛(**bull**) <livestock|牲畜>；太陽隊 **(Phoenix Suns)** ⇨太陽(**Sun**) <celestial|天體>；花木蘭(**Mulan**)⇨木蘭(**magnolia**)<FlowerGrass|花草>	27% (27/100)
10	Unknown word problem	4% (4/100)
11	Error of word segmentation	2% (2/100)

* Types 1 to 3 are contrast to the confirming principles of permissible NVEF knowledge mentioned in section 3.2, respectively.

** Type 4 contents principles (a), (b), and (c) in section 3.2 but there is no proper combination in that sentence.

Table 8. Examples of the eleven types of non-permissible NVEF knowledge. (The English words in parentheses, symbols [] and <> are there for explanatory purposes only.)

NP type	Sentence (English explanation)	**Noun** (English explanation) **DEF**	**Verb** (English explanation) **DEF**
1	警方維護地方[治安]<辛勞> (Police work hard to safeguard the locality security.)	治安 (public security) attribute|屬性,circumstances|境況,safe|安,politics|政,&organization|組織	辛勞 (work hard) endeavour|賣力
2	<模糊>的[白宮]景象 (White House looked vague in the heavy fog.)	白宮 (White House) house|房屋,institution|機構,#politics|政,(US|美國)	模糊 (vague) PolysemousWord|多義詞,CauseToDo|使動,mix|混合
3	<生活>條件[不足] (Lack of living condtions)	不足 (lackness) attribute|屬性,fullness|空滿,incomplete|缺,&entity|實體	生活 (life) alive|活著
4	網路帶給[企業]許多<便利> (Internet brings numerous benefits to industries.)	企業 (Industry) InstitutePlace|場所,*produce|製造,*sell|賣,industrial|工,commercial|商	便利 (benefit) benefit|便利
5	<盈盈>[笑靨] (smile radiantly)	笑靨 (a smiling face) part|部件,%human|人,skin|皮	盈盈 (an adjective, normally to describe a beauty's smile) exist|存在
6	保費較貴的<壽險>[保單] (higher fare life insurance policy)	保單 (insurance policy) bill|票據,*guarantee|保證	壽險 (life insurance) guarantee|保證,scope=die|死,commercial|商
7	債券型基金吸金[存款]<失血> Bond foundation makes profit but savings is loss	存款 (bank savings) money|貨幣,$SetAside|留存	失血 (bleed or loss(only use in finance diction)) bleed|出血
8	華南[銀行] 中山<分行> (Hwa-Nan Bank Jung-San Branch)	銀行 (bank) InstitutePlace|場所,@Set Aside|留存,@TakeBack|取回,@lend|借出,#wealth|錢財,commercial|商	分行 (branch) separate|分離
9	[根據]<調查> (according to the investigation)	根據 (evidence) information|信息	調查 (investigate) investigate|調查
10	<零售>[通路] (retail sell routes)	通路 (route) facilities|設施,route|路	零售 (retail sell) sell|賣
11	從今日<起到> 5[月底] (from today to the end of May)	月底 (the end of month) time|時間,ending|末,month|月	起到 (to elaborate) do|做

Table 8 gives the examples for the eleven types of NP-NVEF knowledge. From Tables 8 and 9, 11% of NP-NVEF cases can be resolved by correcting the error lexicon in original Hownet. For the four NLP errors, these cases could be improved with the support of other techniques such as WSD ([Resnik *et al*. 2000, Yang *et al*. 2002]), unknown word identification ([Chang *et al*. 1997, Lai *et al*. 2000, Chen *et al*. 2002, Sun *et al*. 2002 and Tsai *et al*. 2003]) and word segmentation ([Sproat *et al*. 1996, Teahan *et al*. 2000]).

4. Conclusion and Directions for Future Research

In this paper, we present an auto-discovery system of NVEF knowledge that can be used to automatically generate a large scale NVEF knowledge for NLP. The experimental results shows

that AUTO-NVEF achieves a NVEF accuracy of 98.52%. By applying AUTO-NVEF to the 2001 *UDN* corpus, we create 167,203 NVEF sense-pairs (8.6M) and 317,820 NVEF word-pairs (10.1M) on the NVEF-KR tree. Using this collection of NVEF knowledge, we have designed an NVEF word-pair identifier [Tsai *et al*. 2002] to achieve a WSD accuracy of 93.7% and a STW accuracy of 99.66% for the NVEF related portion in Chinese sentences. The acquired NVEF knowledge can cover 48% and 50% of NV-sentences in ***ASBC*** and in 2001 ***UDN*** corpus, respectively.

Our database for the NVEF knowledge has not been completed. Currently, there are 66.34% (=6,641/10,011) of the noun-senses in Hownet have been considered in the NVEF knowledge construction. The remaining 33.66% of the noun-senses in Hownet not dealt with yet are caused by two problems: (1) those words with ploy-noun-senses or poly-verb-senses, which are difficult to be resolved by WSD, especially those single-character words; and (2) corpus sparseness. We will continue expanding our NVEF knowledge through other corpora. The mechanism of AUTO-NVEF will be extended to auto-generate other meaningful co-occurrence semantic restrictions, in particular, noun-noun association frame (NNAF) pairs, noun-adjective grammar frame (NAGF) pairs and verb-adverb grammar frame (VDGF) pairs. As of our knowledge, the NVEF/NNAF/NAGF/VDGF pairs are the four most important co-occurrence semantic restrictions for language understanding.

Since the creation of NVEF knowledge bears no particular application in mind, and still it can be used to effectively resolve the WSD and STW problems, the NVEF knowledge is potentially application independent for NLP. We shall further investigate the effectiveness of NVEF knowledge in other NLP applications, such as syllable/speech understanding and full/shallow parsing.

5. Acknowledgements

We are grateful to our colleagues in the Intelligent Agent Systems Lab. (IASL), Li-Yeng Chiu, Mark Shia, Gladys Hsieh, Masia Yu, Yi-Fan Chang, Jeng-Woei Su and Win-wei Mai, who helped us create and verify all the necessary NVEF knowledge and tools for this study. We would also like to thank Prof. Zhen-Dong Dong for providing us with the Hownet dictionary.

Reference

Carey, S., "The origin and evolution of everyday concepts (In R. N. Giere, ed.)," *Cognitive Models of Science*, Minneapolis: University of Minnesota Press, 1992.

Chang, J. S. and K. Y. Su, "An Unsupervised Iterative Method for Chinese New Lexicon Extraction," *International Journal of Computational Linguistics & Chinese language Processing*,

1997Choueka, Y. and S. Lusignan, "A Connectionist Scheme for Modeling Word Sense Disambiguation," *Cognition and Brain Theory*, 6 (1) 1983, pp.89-120.

Chen, K. J. and W. Y. Ma, "Unknown Word Extraction for Chinese Documents," *Proceedings of 19th COLING 2002*, Taipei, pp.169-175

Chu, S. C. R., *Chinese Grammar and English Grammar: a Comparative Study*, The Commerical Press, Ltd. The Republic of China, 1982

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

Dong, Z. and Q. Dong, *Hownet*, http://www.keenage.com/

Fromkin, V. and R. Rodman, *An Introduction to Language*, Sixth Edition, Holt, Rinehart and Winston, 1998

Krovetz, R. and W. B. Croft, "Lexical Ambiguity and Information Retrieval," *ACM Transactions on Information Systems*, 10 (2), 1992, pp.115-141.

Lai, Y. S. and Wu, C. H., "Unknown Word and Phrase Extraction Using a Phrase-Like-Unit-based Likelihood Ratio," *International Journal of Computer Processing Oriental Language*, 13(1), pp.83-95

Li, N. C. and S. A. Thompson, Mandarin Chinese: a Functional Reference Grammar, The Crane Publishing Co., Ltd. Taipei, Taiwan, 1997

On-Line United Daily News, http://udnnews.com/NEWS/

Resnik, P. and D. Yarowsky, "Distinguishing Systems and Distinguishing Senses: New Evaluation Methods for Word Sense Disambiguation," *Natural Language Engineering*, 5 (3), 2000, pp.113-133.

Small, S., and G. Cottrell, and M. E. Tannenhaus, *Lexical Ambiguity Resolution*, Morgan Kaufmann, Palo Alto, Calif., 1988.

Sun, J., J. Gao, L. Zhang, M. Zhou and C. Huang, "Chinese Named Entity Identification Using Class-based Language Model," *Proceedings of 19th COLING 2002*, Taipei, pp.967-973

Sproat, R. and C. Shih, "A Stochastic Finite-State Word-Segmentation Algorithm for Chinese," *Computational Linguistics*, 22(3), 1996, pp.377-404

Teahan, W.J., Wen, Y., McNab, R.J., Witten, I.H., "A compression-based algorithm for chinese word segmentation," *Computational Linguistics*, 26, 2000, pp.375-393

Tsai, J. L, W. L. Hsu and J. W. Su, "Word sense disambiguation and sense-based NV event-frame identifier," *Computational Linguistics and Chinese Language Processing,* Vol. 7, No. 1, February 2002, pp.29-46

Tsai, J. L, W. L. Hsu, "Applying NVEF Word-Pair Identifier to the Chinese Syllable-to-Word Conversion Problem," *Proceedings of 19th COLING 2002*, Taipei, pp.1016-1022

Tsai, J. L, C. L. Sung and W. L. Hsu, "Chinese Word Auto-Confirming Agent," *Proceeding of ROCLING XV*, 2003

Yang, X. and Li T., “A study of Semantic Disambiguation Based on HowNet,” *Computational Linguistics and Chinese Language Processing,* Vol. 7, No. 1, February 2002, pp.47-78

陳克健，洪偉美，“中文裏「動—名」述賓結構與「動—名」偏正結構的分析，” *Communication of COLIPS*, 6(2), 1996, pp.73-79

胡裕樹，范曉，*動詞研究*，河南大學出版社，1995

Appendix A. A Sample Table of the Main Features of Nouns and their 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. Example Mappings of FPOS and NV Word-Pairs

FPOS	NV word-pairs	Example, [] stands for noun and <> stands for verb
$N_1V_2ADJ_3N_4$	N_1V_2 & N_4V_2	[學生]<購買>許多[筆記本]
N_1V_2	N_1V_2	[雜草]<枯萎>
N_1 ADJ_2 ADV_3V_4	N_1V_4	[意願]遲未<回升>

Appendix C. Ten Examples of General-Keeping (GK) Conditions

Noun-sense class	Verb DEF	Example, [] stands for noun and <> stands for verb
微生物(bacteria)	own|有	已經使[細菌]<具有>高度抗藥性
位置類(location)	arrive|到達	若正好<蒞臨>[西班牙]
植物類(plant)	decline|衰敗	田中[雜草]<枯萎>
人工物(artifact)	buy|買	民眾不需要急著<購買>[米酒]
天然物(natural)	LeaveFor|前往	立刻驅船<前往>蘭嶼[海域]試竿
事件類(event)	alter|改變	批評這會<扭曲>[貿易]
精神類(mental)	BecomeMore|增多	民間投資[意願]遲未<回升>
現象類(phenomena)	announce|發表	做任何<公開>[承諾]
物形類(Shape)	be|是,all|全	由於從腰部以下<都是>合身[線條]
地點類(place)	from|相距	<距離>[小學]七百公尺