Auto-Generation of NVEF Knowledge 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

Noun-verb event frame (NVEF) knowledge in conjunction with an NVEF word-pair identifier [Tsai *et al*. 2002] comprises a system to support natural language processing (NLP) and natural language understanding (NLU). In [Tsai *et al*. 2002a], we demonstrated that NVEF knowledge can be used effectively to resolve the Chinese word-sense disambiguation (WSD) problem with 93.7% accuracy for nouns and verbs. In [Tsai *et al*. 2002b], we have shown that NVEF knowledge can be applied on the Chinese syllable-to-word (STW) conversion problem to achieve 99.66% accuracy for the NVEF related portions of Chinese sentences. In [Tsai *et al*. 2002a], we defined a collection of NVEF knowledge as an NVEF word-pair (a meaningful NV word-pair) and its corresponding NVEF sense-pairs. No existing methods exist that can fully automatically find collections of NVEF knowledge from Chinese sentences. We propose a method for automatically acquiring large-scale NVEF knowledge without human intervention in order to identify a large, varied range of NVEF-sentences (sentences containing at least one NVEF word-pair). The auto-generation of NVEF knowledge (AUTO-NVEF) includes four major processes: (1) segmentation check; (2) Initial Part-of-Speech (IPOS) sequence generation; (3) NV knowledge generation; and (4) NVEF knowledge auto-confirmation.

Our experiment results show that AUTO-NVEF achieves 98.52% accuracy for news and 96.41% for specific text types, which include research report, classical literature and modern literature. AUTO-NVEF automatically discovered over 400,000 NVEF word-pairs from the 2001 *United Daily News* (2001 *UDN*) corpus. According to our estimation, the acquired NVEF knowledge from 2001 *UDN* helps to identify 54% of NVEF-sentences in the *Academia Sinica Balanced Corpus* (*ASBC*), and 60% in the 2001 *UDN* corpus.

We plan to expand NVEF knowledge to be able to identify more than 75% of NVEF-sentences in *ASBC*. We will also apply the acquired NVEF knowledge to support other NLP and NLU researches, such as machine translation, shallow parsing, syllable and speech understanding and text indexing. The auto-generation of bilingual, especially Chinese-English, NVEF knowledge will be also addressed in our future work.

Keywords: natural language understanding, verb-noun collection, machine learning, HowNet

1. Introduction

The most challenging problem in natural language processing (NLP) is programming computers to understand natural languages. For humans, efficient syllable-to-word (STW) conversion and word sense disambiguation (WSD) arise naturally when a sentence is understood. In designing a natural language understanding (NLU) system, methods that enable consistent STW and WSD are critical yet difficult to attain. 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 multiple senses, to understand language, efficient word sense disambiguation (WSD) becomes a critical issue for NLU systems. As per 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). That is, the relationship between a word and each of the others in the sentence can effectively resolve ambiguity. From [Small *et al*. 1988; Krovetz *et al*. 1992; Resnik *et al*. 2000], most ambiguities occur with nouns and verbs. Object-event (i.e. noun-verb) distinction is the most prominent ontological distinction for humans [Carey 1992]. Tsai *et al*. (2002a) have shown that the knowledge of meaningful noun-verb (NV) word-pairs and their corresponding sense-pairs in conjunction with an NVEF word-pair identifier can achieve a WSD accuracy of 93.7% for NV-sentences (sentences that contain at least one noun and one verb).

According to [胡裕樹 *et al*. 1995; 陳克健 *et al*. 1996; Fromkin *et al*. 1998; 朱曉亞 2001; 陳昌來 2002; 劉順 2003], the most important content word relationship in sentences is noun-verb construction. For most languages, subject-predicate (SP) and verb-object (VO) are the two most common NV constructions (or meaningful NV word-pairs). In Chinese, SP and VO constructions can be found in three language units: compounds, phrases and sentences [Li *et al*. 1997]. Modifier-head (MH) and verb-complement (VC) are another two meaningful NV word-pairs which are only found in phrases and compounds. Consider the meaningful NV word-pair *汽車-進口*(**car**, **import**). It is a MH construction in the Chinese compound 進口汽車(import car) and is a VO construction in the Chinese phrase *進口*許多*汽車*(**import** many **cars**). In [Tsai *et al*. 2002a], we called a meaningful NV word-pair a *noun-verb event frame* (*NVEF*) word-pair. Combining the NV word-pair *汽車-進口* and its sense-pair **Car-Import** is a *collection* of NVEF knowledge. Since a complete event frame usually contains a predicate and its arguments, an NVEF word-pair can be a full or a partial event frame construction.

In Chinese, syllable-to-word entry is the most popular input method. Since the average number of characters sharing the same phoneme is 17, efficient STW conversion becomes an indispensable tool. In [Tsai *et al*. 2002b], we have shown that NVEF knowledge can be used to achieve an STW accuracy of 99.66% for converting NVEF related words in Chinese. We created a semi-automatic generation of NVEF knowledge method in [Tsai *et al*. 2002a]. This method

uses the NV frequencies in sentences groups to generate NVEF candidates to be filtered by human editors. This becomes labor-intensive when creating a large-scale NVEF knowledge. To our knowledge, no existing methods that can fully auto-extract a large number of NVEF knowledge from Chinese text. In literatures, most methods of auto-extracting Verb-Noun collections (i.e. meaningful NV word-pairs) are focus on English and achieve 60% to 65% precision and 70% to 75% recall [Benson *et al*. 1986; Church *et al*. 1990; Smadja 1993; Smadja *et al*. 1996; Lin 1998; Huang *et al*. 2000; Jian 2003]. However, the issue of VN collections is focus on extracting meaningful NV word-pairs not on NVEF knowledge. In this paper, we propose a new method that *automatically* generates NVEF knowledge from running texts and constructs large-scale NVEF knowledge.

This paper is arranged as follows. In Section 2, we detail the auto-generation of NVEF knowledge. Experiment results and analyses are given in Section 3. The conclusion is given and future research ideas are discussed in Section 4.

2. Development of NVEF Knowledge Auto-Generation

For our auto-generate NVEF knowledge (AUTO-NVEF) system, we use HowNet 1.0 [Dong 1999] as a system dictionary. This system dictionary provides 58,541 Chinese words and their corresponding parts-of-speech (POS) and word senses (called DEF in HowNet). Contained in this dictionary are 33,264 nouns and 16,723 verbs, as well as 16,469 senses comprised of 10,011 noun-senses and 4,462 verb-senses.

From 1999, HowNet has become one of famous Chinese-English bilingual knowledge-base dictionaries for Chinese NLP research fields. Machine translation (MT) is one of typical applications of HowNet. The interesting issues of (1) overall picture of HowNet, (2) comparison between HowNet [Dong 1999], WordNet [Miller 1990; Fellbaum 1998], Suggested Upper Merged Ontology (SUMO) [Niles *et al*. 2001; Subrata *et al*. 2002; Chung *et al*. 2003] and VerbNet [Dang *et al*. 2000; Kipper *et al*. 2000] and (3) typical applications of HowNet can be found in the 2nd tutorial of *IJCNLP-04* [Dong 2004].

2.1 Definition of NVEF Knowledge

The sense of a word is defined as its definition of concept (DEF) in HowNet. Table 1 lists three different senses of the Chinese word 車(Che[surname]/car/turn). In HowNet, the DEF of a word consists of its main feature and all secondary features. For example, in the DEF "character|文字,surname|姓,human|人,ProperName|專" of the word 車(Che[surname]), 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 inherits features in the

hypernym-hyponym hierarchy. There are approximately 1,500 of these features in HowNet. Each is called a *sememe*, which refers to the smallest semantic unit that cannot be reduced.

Table 1. The three different senses of the Chinese word 車(Che[surname]/car/turn)

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

[a] C.Word means Chinese word; E.Word means English word

As we mentioned, a meaningful NV word-pair is a noun-verb event-frame word-pair (*NVEF word-pair*), such as 車 - 行駛(Che[surname]/car/turn, move). In sentences, an NVEF word-pair can be of SP or VO construction; in phrases/compounds, an NVEF word-pair can be of SP, VO, MH or VC construction. From Table 1, the only meaningful NV sense-pair for 車 - 行駛(car, move) is **LandVehicle\|車 - VehicleGo\|駛**. Here, the combination of the NVEF sense-pair **LandVehicle\|車 - VehicleGo\|駛** and the NVEF word-pair 車 - 行駛 is a *collection* of NVEF knowledge.

2.2 Knowledge Representation Tree for NVEF Knowledge

To effectively represent NVEF knowledge, we have proposed an NVEF knowledge representation tree (NVEF KR-tree) to store, edit and browse acquired NVEF knowledge. The details of the NVEF KR-tree shown below are taken from [Tsai *et al*. 2002a].

The two types of nodes in the KR-tree are *function nodes* and *concept nodes*. Concept nodes refer to words and senses (DEF) of NVEF knowledge. Function nodes define the relationships between the parent and children concept nodes. According to each main feature of noun senses in HowNet, we classify these noun senses into fifteen subclasses. These subclasses 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 table of the fifteen main noun features in each noun-sense subclass.

As shown in Figure 1, the three function nodes that construct the collection of NVEF knowledge (LandVehicle\|車- VehcileGo\|駛) are:

(1) **Major Event** (主要事件): The content of the major event 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 VehicleGo\|駛 in Figure 1. To describe various relationships between noun-sense and verb-sense sub-

classes, we design three subclass sense-symbols: = means *exact*, & means *like*, and % means *inclusive*. For example, 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, S_3/W_3 is in the *exact*-subclass of =LandVehicle|車,police|警; S_1/W_1 and S_2/W_2 are in the *like*-subclass of &LandVehicle|車,*transport|運送; and S_1/W_1, S_2/W_2, and S_3/W_3 are in the *inclusive*-subclass of %LandVehicle|車.

(2) **Word Instance** (實例): The content of word instance children is words belonging to the sense subclass of its parent node. These words are self-learned through the sentences under the Test-Sentence nodes.

(3) **Test Sentence** (測試題): The content of test sentence children is the selected test NV-sentence that provides language context for its corresponding NVEF knowledge.
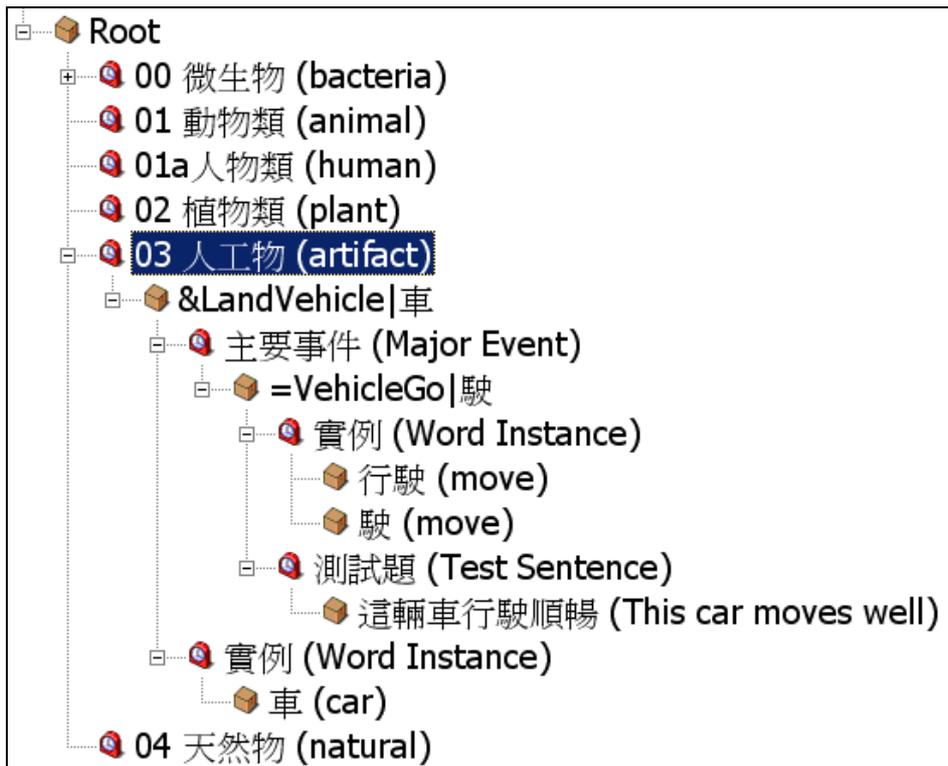

Figure 1. An illustration of the KR-tree using 人工物 (artifact) as an example noun-sense subclass. The English words in parentheses are provided for explanatory purposes only.

2.3 Auto-Generate NVEF Knowledge

AUTO-NVEF automatically discovers meaningful NVEF sense/word-pairs (NVEF knowledge) from Chinese sentences. Figure 2 is the AUTO-NVEF flow chart. There are four major processes in AUTO-NVEF; their details are shown in Figure 2, and Table 2 gives a step by step example. A detailed description of each process is as follows.
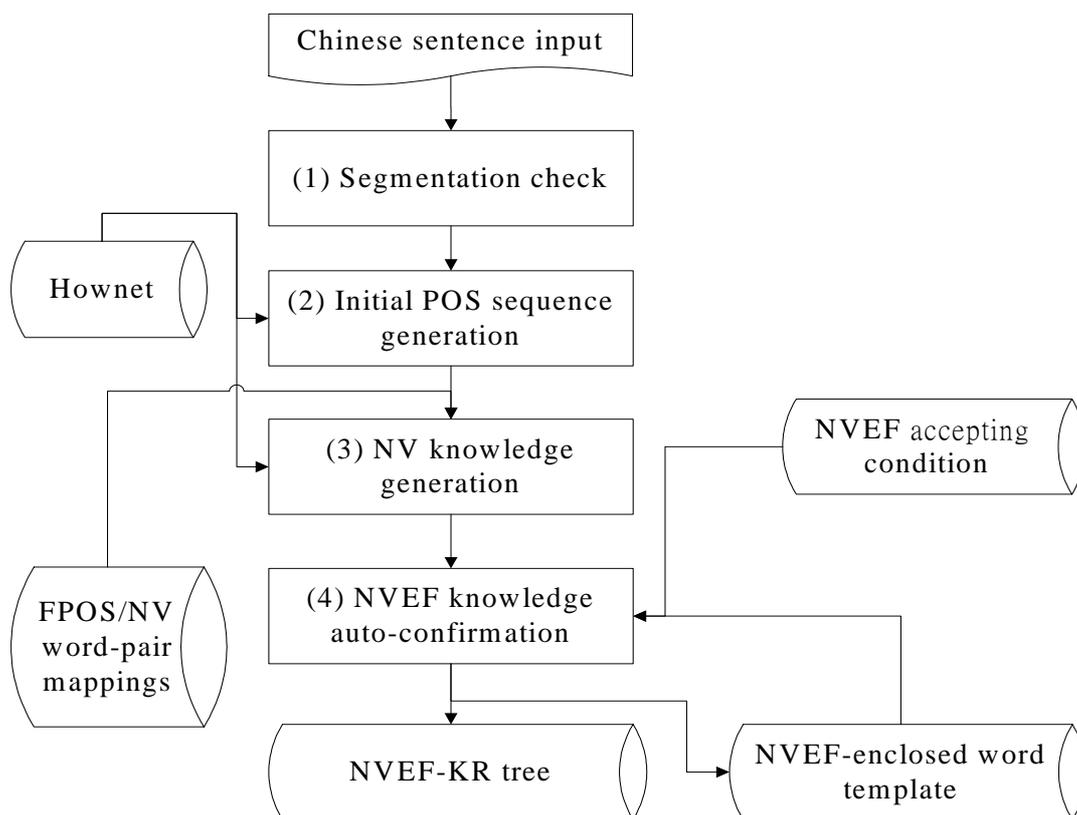
```
         Chinese sentence input
                   │
                   ▼
        ┌────────────────────────┐
        │ (1) Segmentation check  │
        └────────────────────────┘
                   │
  ┌─────────┐      ▼
  │ Hownet  │──► ┌────────────────────────┐
  └─────────┘    │ (2) Initial POS sequence│
                 │      generation         │
                 └────────────────────────┘
                   │
                   ▼                         ┌──────────────────────┐
              ┌────────────────────────┐     │ NVEF accepting       │
          ──► │ (3) NV knowledge        │     │    condition         │
              │      generation         │     └──────────────────────┘
              └────────────────────────┘
                   │
                   ▼
  ┌──────────┐  ┌────────────────────────┐
  │ FPOS/NV  │  │ (4) NVEF knowledge      │
  │ word-pair│  │   auto-confirmation     │◄──
  │ mappings │  └────────────────────────┘
  └──────────┘     │
                   ▼
         ┌──────────────────┐   ┌──────────────────────┐
         │  NVEF-KR tree    │──►│ NVEF-enclosed word    │
         └──────────────────┘   │     template          │
                                └──────────────────────┘
```

Figure 2. AUTO-NVEF flow chart

Process 1. Segmentation check: In this stage, a Chinese sentence is segmented by two strategies: *forward (left-to-right) longest word first*, and *backward (left-to-right) longest word first*. From [Chen *et al*. 1986], the "longest syllabic word first strategy" is effective for Chinese word segmentation. If both forward and backward segmentations are equal (forward=backward) and the word number of the segmentation is greater than one, this segmentation result will be sent to **process 2**; otherwise, a *NULL* segmentation will be sent. Table 3 is a comparison of word-segmentation accuracies for forward, backward and forward=backward strategies using the *Chinese Knowledge Information Processing* (CKIP) lexicon [CKIP 1995]. The word segmentation accuracy is the ratio of correct segmented sentences to all sentences of the *Academia Sinica Balancing Corpus* (**ASBC**) [CKIP 1996]. A correct segmented sentence means the seg-

6

mented result exactly matches its corresponding segmentation in **ASBC**. Table 3 shows that the technique of forward=backward achieves the best word segmentation accuracy.

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 audience members 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 members)
(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)	NV1 = 現場/place|地方,#fact|事情/N
	- 湧入(yong3 ru4)/GoInto|進入/V
	NV2 = 觀眾/human|人,*look|看,#entertainment|藝,#sport|體育,*recreation|娛樂/N
	- 湧入(yong3 ru4)/GoInto|進入/V
(4)	NV1 is NVEF knowledge by NVEF accepting-condition; learned NVEF template is [音樂會 NV 許多]
	NV2 is NVEF knowledge by NVEF accepting-condition; learned NVEF template is [現場 V 許多 N]

Table 3. A comparison of word-segmentation accuracies for backward, forward and backward = forward strategies. Test sentences are from **ASBC** and the dictionary is CKIP lexicon.

	Backward	Forward	Backward = Forward
Accuracy	82.5%	81.7%	86.86%
Recall	100%	100%	89.33%

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

1) For segmentation result $w_1/w_2/\ldots/w_{n-1}/w_n$ from *process 1*, our algorithm computes the POS of w_i, where i = 2 to n. Then, it computes the following two sets: a) the *following POS/frequency set* of w_{i-1} according to the **ASBC** and b) the *HowNet POS set* of w_i. It then computes the POS intersection of the two sets. Finally, it selects the POS with the largest frequency in the POS intersection for the POS of w_i. If there is zero or more than one POS with the largest frequency, the POS of w_i will be set to *NULL* POS.

2) For the POS of w_1, it selects 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 from first two steps, it then generates the ***initial POS sequence*** (***IPOS***). Take the Chinese segmentation 生/了 as an example. The following POS/frequency set of the Chinese word 生(to bear) is {N/103, PREP/42, STRU/36, V/35, ADV/16, CONJ/10, ECHO/9, ADJ/1}(see Table 4 for tags defined in HowNet). The HowNet POS set of the Chinese word 了(a Chinese satisfaction indicator) 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 tags and HowNet POS tags.

Table 4. A mapping list of CKIP POS tags and HowNet POS tags

	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: This process will be triggered if the *IPOS* output of *process 2* does not include any *NULL* POS. The steps of this process are given as follows.

1) Compute the *final POS sequence* (*FPOS*). This step translates an *IPOS* into a *FPOS*. For each continuous nouns sequence of *IPOS*, the last noun will be kept and the other nouns will be dropped. It is because a contiguous nouns sequence in Chinese is usually a compound and its head is the last noun. 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) and its IPOS $N_1N_2V_3ADJ_4N_5$ as an example. Since it has a continuous nouns sequence音樂會(N_1)現場(N_2), the *IPOS* will be translated into *FPOS* $N_1V_2ADJ_3N_4$, where N_1=現場, V_2=湧入, ADJ_3=許多and N_4=觀眾.

2) Generate NV word-pairs. According to the *FPOS* mappings and their corresponding NV word-pairs (see Appendix B), AUTO-NVEF generates the NV word-pairs. In this study, we created more than one hundred *FPOS* mappings and their corresponding NV word-pairs. Consider the above mentioned *FPOS* $N_1V_2ADJ_3N_4$, where N_1=現場, V_2=湧入, ADJ_3=許多 and N_4=觀眾. Since the corresponding NV word-pairs for the *FPOS* $N_1V_2ADJ_3N_4$ are N_1V_2 and N_4V_2, AUTO-NVEF will generate two NV word-pairs 現場**(N)**湧入**(V)** and湧入**(V)**觀眾**(N)**. In [朱曉亞 2001], there are some useful semantic structure patterns of Modern Chinese sentences for creating the *FPOS* mappings and their corresponding NV word-pairs.

3) Generate NV knowledge. According to HowNet, AUTO-NVEF computes all NV sense-pairs for the generated NV word-pairs. Consider the generated NV word-pairs 現場**(N)**湧入**(V)** and 湧入**(V)**觀眾**(N).** AUTO-NVEF will generate two collections of NV knowledge:

   NV1 = [現場(locale)/place|地方,#fact|事情/N] - [湧入(enter)/GoInto|進入/V], and

   NV2 = [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, AUTO-NVEF automati-

cally confirms whether the generated NV knowledge is or is not NVEF knowledge. The two auto-confirmation procedures are below.

(a) **NVEF accepting condition (NVEF-AC) check**: Each NVEF accepting condition is constructed by a noun-sense class (such as 人物類[human]) defined in [Tsai *et al*. 2002a] and a verb main feature (such as GoInto|進入) defined in HowNet [Dong 1999]. In [Tsai *et al*. 2002b], we created 4,670 NVEF accepting conditions from manually confirmed NVEF knowledge. In this procedure, if the noun-sense class and the verb main feature of the generated NV knowledge can fit at least one NVEF accepting condition, the generated NV knowledge will be auto-confirmed as NVEF knowledge and will be sent to NVEF KR-tree. Appendix C gives the ten NVEF accepting 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. An NVEF-EW template is composed of all left words and right words of an 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 NVEF-EW templates are auto-generated from: 1) the collection of manually confirmed NVEF knowledge in [Tsai *et al*. 2002], 2) the on-line collection of automatically confirmed NVEF knowledge by AUTO-NVEF and 3) the manually created NVEF-EW templates. In this procedure, if the NVEF-EW template of a generated NV word-pair matches at least one NVEF-EW template, the NV knowledge will be auto-confirmed as NVEF knowledge.

3. Experiments

To evaluate the performance of the proposed auto-generation of NVEF knowledge, we define the NVEF accuracy and NVEF-identified sentence ratio by Equations (1) and (2), respectively.

NVEF accuracy =
  # of meaningful NVEF knowledge / # of total generated NVEF knowledge.　　　　(1)
NVEF-identified sentence ratio =
  # of NVEF-identified sentences / # of total NVEF-sentences.　　　　(2)

In Equation (1), a meaningful NVEF knowledge means that the generated NVEF knowledge has been manually confirmed as a collection of NVEF knowledge. In Equation (2), if a Chinese sentence can be identified with at least one NVEF word-pair by the generated NVEF knowledge in conjunction with the NVEF word-pair identifier in [Tsai *et al*. 2002a], this sen-

tence is called an **NVEF-identified sentence**. If a Chinese sentence contains at least one NVEF word-pairs, it is called an **NVEF-sentence**. We estimate that about 70% of the Chinese sentences in *ASBC* are NVEF-sentences.

3.1 User Interface for Manually Confirming NVEF Knowledge

A user interface that manually confirming generated NVEF knowledge is as shown in Figure 3. With it, evaluators (native Chinese speakers) can review the generated NVEF knowledge and determine whether or not it is meaningful NVEF knowledge. Take the Chinese sentence 高度壓力(High pressure)使(make)有些(some)人(people)食量(eating capacity)減少 (decrease) as an example. AUTO-NVEF will generate an NVEF knowledge collecting that includes 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)]. The confirmation principles of meaningful NVEF knowledge are given as follows. Appendix D is a snapshot to demonstrate the designed user interface for evaluators for manually confirming generated NVEF knowledge.

Chinese sentence	高度壓力(High pressure)使(make)有些(some)人(people)食量(eating capacity)減少 (decrease)		
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 user interface of NVEF knowledge using the generated NVEF knowledge for the Chinese sentence 高度壓力(High pressure)使(makes)有些(some)人 (people)食量(eating capacity)減少(decrease). The English words in parentheses are for explanatory purposes only. [] indicates nouns and <> indicates verbs.

3.2 Confirmation Principles of Meaningful NVEF Knowledge

Auto-generated NVEF knowledge should be confirmed as meaningful NVEF knowledge if it fits all three of the following principles.

Principle 1. The NV word-pair makes correct POS tags for the given Chinese sentence.
Principle 2. The NV sense-pair and the NV word-pair make sense.
Principle 3. Most inherited NV word-pairs of the NV sense-pair satisfy Principles 1 and 2.

3.3 Experiment Results

For our experiment, we use two corpora. One is the 2001 *UDN* 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. The other is the collection of spe-

cific text types which include research reports, classical literature and modern literature. The details of training, testing corpora and test sentence sets are given below.

(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 4,670 NVEF accepting conditions.

(2) **Testing corpora.** One corpus is the collection of Chinese sentences extracted from the 2001 *UDN* corpus from October 1, 2001 to December 31, 2001. The other is the collection of specific text types which include research reports, classical literature and modern literature.

(3) **Test sentence sets**. From the first testing corpus, we randomly select all sentences extracted from the news of October 27, 2001, November 23, 2001 and December 17, 2001 on 2001 *UDN* to be our first test sentence set. From the second testing corpus, we select a research report, a classical novel and a modern novel for our second test sentences set.

All NVEF knowledge acquired by AUTO-NVEF from the testing corpora was manually confirmed by evaluators. Tables 5a and 5b show the experiment results. These tables show that our AUTO-NVEF achieves 98.52% NVEF accuracy for news and 96.41% for specific text types.

Table 5a. Experiment results of AUTO-NVEF for news

News article date	NVEF accuracy		
	NVEF-AC	NVEF-EW	NVEF-AC + NVEF-EW
October 27, 2001	99.54%(656/659)	98.43%(439/446)	99.10% (1,095/1,105)
November 23, 2001	98.75%(711/720)	95.95%(379/395)	97.76% (1,090/1,115)
December 17, 2001	98.74%(1,015/1,028)	98.53%(1,141/1,158)	98.63% (2,156/2,186)
Total Average	98.96%(2,382/2,407)	98.00%(1,959/1,999)	98.52% (4,341/4,406)

Table 5b. Experiment results of AUTO-NVEF for specific text types

Text type	NVEF accuracy		
	NVEF-AC	NVEF-EW	NVEF-AC + NVEF-EW
Technique Report	97.12%(236/243)	96.61%(228/236)	96.86% (464/479)
Classic novel	98.64%(218/221)	93.55%(261/279)	95.80% (479/500)
Modern novel	98.18%(377/384)	95.42%(562/589)	96.51% (939/973)
Total Average	98.00%(831/848)	95.20%(1,051/1,104)	96.41% (1,882/1,952)

When we apply AUTO-NVEF to the entire 2001 *UDN* corpus, it auto-generates 173,744 NVEF sense-pairs (8.8M) and 430,707 NVEF word-pairs (14.1M). Within this data, 51% is generated through NVEF accepting conditions (human-editing knowledge) and 49% is gener-

ated through NVEF-enclosed word templates (machine-learning knowledge). From Table 5a and 5b, it shows that the average accuracies of NVEF knowledge generated by NVEF-AC and NVEF-EW for news and specific texts achieve 98.71% and 97.00%, respectively. This result indicates that our AUTO-NVEF has the ability to simultaneously maintain its precision and extend NVEF-EW knowledge, like as snowball effect, to generate large-scale NVEF knowledge without human intervention. The results also support that the best method to overcome *Precision-Recall Tradeoff* problem for NLP is based on linguistic knowledge and statistical constraints, i.e. hybrid approach [Huang *et al.* 1996; Tsai *et al.* 2003].

3.3.1 Analysis and Classification of NVEF Knowledge

From noun and verb positions of NVEF word-pairs in Chinese sentences, NVEF knowledge can be classified into four NV-position types: **N:V**, **N-V**, **V:N** and **V-N**, where **:** means next to and - means nearby. Table 6a shows examples and the percentages of the four NV-position types of generated NVEF knowledge. The ratios (percentages) of the collections of **N:V**, **N-V**, **V:N** and **V-N** are 12.41%, 43.83% 19.61% and 24.15%, respectively. Table 6a shows that an NVEF word-pair, such as 工程-完成(**Construction**, **Complete**), can be an **N:V**, **N-V**, **V:N** or **V-N** in sentences. To our generated NVEF knowledge, the maximum and average character numbers between noun and verb of generated NVEF knowledge are 27 and 3, respectively.

Table 6a. An illustration of four NV-position types of NVEF knowledge and their ratios. The English words in parentheses are for explanatory purposes only. [] indicates nouns and <> indicates verbs.

Type	Example Sentence	Noun / DEF	Verb / DEF	Percentage
N:V	[*工程*]<*完成*> (The construction is now completed)	工程 (construction) affairs|事務,industrial|工	完成 (complete) fulfill|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) fulfill|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) fulfill|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) fulfill|實現	12.41%

From the character numbers of noun and verb of NVEF word-pairs, we classify NVEF

knowledge into four NV-word-length types: **N1V1**, **N1V2+**, **N2+V1** and **N2+V2+**, where N1 and V1 means single-character nouns and verbs; N2+ and V2+ means multi-character nouns and verbs. Table 6b shows examples and the percentages of the four NV-word-length types of a manually created NVEF knowledge for randomly selected 1,000 *ASBC* sentences. From the manually created NVEF knowledge, we estimate that the ratios of the collections of **N1V1**, **N1V2+**, **N2+V1** and **N2+V2+** NVEF word-pairs are 6.4%, 6.8%, 22.2% and 64.6%, respectively. According to this NVEF knowledge, we estimate that the auto-generated NVEF Knowledge (for 2001 *UDN*) in conjunction with the NVEF word-pair identifier [Tsai *et al.* 2002] can identify 54% of NVEF-sentences in *ASBC*.

Table 6b. An illustration of four NV-word-length types of manually-edited NVEF knowledge for randomly selected 1,000 *ASBC* sentences and their ratios. The English words in parentheses are for explanatory purposes only. [] indicates nouns and <> indicates verbs.

Type	Example Sentence	Noun	Verb	Percentage
N1V1	然後就<*棄*>[*我*]而去	我(I)	棄(give up)	6.4%
N1V2+	<*覺得*>[*他*]很孝順	他(he)	覺得(feel)	6.8%
N2+V1	<*買*>了[*可樂*]來喝	可樂(cola)	買(buy)	22.2%
N2+V2+	<*引爆*>另一場美西[*戰爭*]	戰爭(war)	引爆(cause)	64.6%

Table 6c shows the Top 5 single-character verbs of N1V1 and N2+V1 NVEF word-pairs and their ratios. Table 6d shows the Top 5 multi-character verbs of N1V2+ and N2+V2+ NVEF word-pairs and their ratios. From Table 6c, the ratios of N2+是 and N2+有 NVEF word-pairs both are greater than those of other single-character verbs. Thus, the N2+是 and N2+有 NVEF knowledge was worthy to be considered in our AUTO-NVEF. On the other hand, we found that there are 3.2% of NVEF-sentences (or say 2.3% of sentences) are N1V1-only sentences, where N1V1-only sentence means a sentence that only has one N1V1-NVEF word-pair. For example, the Chinese sentence 他(he)說(say)過了(already) is an N1V1-only sentence because it has only one N1V1-NVEF knowledge *他-說*(**he**, **say**). Since (1) N1V1-NVEF knowledge is not critical for our NVEF-based applications and (2) the auto-generation of N1V1 NVEF knowledge is very difficult, the auto-generation of N1V1-NVEF knowledge was not considered in our AUTO-NVEF. In fact, according to the system dictionary, the maximum and average word-sense numbers of single-character are 27 and 2.2; that of multi-character words are 14 and 1.1.

Table 6c. An illustration of Top 5 single-character verbs of **N1V1** and **N2+V1** word-pairs of manually-edited NVEF knowledge for randomly selected 1,000 *ASBC* sentences and their ratios. The English words in parentheses are for explanatory purposes only. [] indicates nouns and <> indicates verbs.

Top	Verb of N1V1 / Example Sentence	Ratio of N1V1	Verb of N2+V1 / Example Sentence	Ratio of N2+V1
1	有(have) / [*我*]<*有*>九項獲參賽資格	16.5%	是(be) / 再來就<*是*>一間陳列樂器的[*房子*]	20.5%
2	是(be) / [*它*]<*是*>做人的根本	8.8%	有(have) / 是不是<*有*>[*問題*]了	15.5%
3	說(speak) / [*他*]<*說*>	7.7%	說(speak) / 而談到成功的秘訣[*妮娜*]<*說*>	3.9%
4	看(see) / <*看*>著[*它*]被卡車載走	4.4%	到(arrive) / 一[*到*]<*陰天*>	3.6%
5	買(buy) / 美國本土的人極少到那兒<*買*>[*地*]	3.3%	讓(let) / <*讓*>現職[*人員*]無處棲身	2.5%

Table 6d. An illustration of Top 5 multi-character verbs of **N1V2+** and **N2+V2+** word-pairs of manually-edited NVEF knowledge for randomly selected 1,000 *ASBC* sentences and their ratios. The English words in parentheses are for explanatory purposes only. [] indicates nouns and <> indicates verbs.

Top	Verb of N1V2+ / Example Sentence	Ratio of N1V2+	Verb of N2+V2+ / Example Sentence	Ratio of N2+V2+
1	吃到(eat) / 你也可能<*吃到*>毒[*魚*]	2.06%	表示(express) / 這位[*官員*]<*表示*>	1.2%
2	知道(know) / [*我*]<*知道*>哦	2.06%	使用(use) / 歌詞<*使用*>日常生活[*語言*]	1.1%
3	喜歡(like) / 至少還有人<*喜歡*>[*他*]	2.06%	沒有(not have) / 我們就<*沒有*>什麼[*利潤*]了	0.9%
4	充滿(fill) / [*心*]裡就<*充滿*>了感動與感恩	2.06%	包括(include) / <*包括*>被監禁的民運[*人士*]	0.8%
5	打算(plan) / [*你*]<*打算*>怎麼試	2.06%	成為(become) / 這種與上司<*成為*>知心[*朋友*]的作法	0.7%

3.3.2 Error Analysis - Non-Meaningful NVEF Knowledge Generated by AUTO-NVEF

One hundred collections of manually confirmed non-meaningful NVEF (NM-NVEF) knowledge from the experiment results are analyzed. We classify these into eleven error types, as shown in Table 7, which lists the NM-NVEF confirmation principles and the ratios for the eleven error types. The first three types comprise 52% of the NM-NVEF cases that do not satisfy NVEF confirmation principles 1, 2 and 3. The fourth type is rare with 1% of the NM-NVEF cases. Types 5, 6 and 7 comprise 11% of the NM-NVEF cases and are caused from incorrect

HowNet lexicon, such as the incorrect DEF (word-sense) *exist*/*存在* for the Chinese word *盈盈* (an adjective, normally used to describe some one's beautiful smile). Types 8, 9, 10 and 11 are referred to as the *four NLP errors* comprising 36% of the NM-NVEF cases. Type 8 is caused by the problem of different word-senses used in Old and Modern Chinese; Type 9 is caused by errors in WSD; type 10 is incurred from the unknown word problem; and Type 11 is caused by incorrect word segmentation.

Table 7. Eleven error types and their confirming principles of non-meaningful NVEF knowledge generated by AUTO-NVEF

Type	Confirming Principles of Non-Meaningful NVEF Knowledge	Percentage
1[*]	NV Word-pair cannot make a correct or sensible POS tag for the Chinese sentence	33% (33/100)
2[*]	The combination of the NV sense-pair (DEF) and the NV word-pair cannot be an NVEF knowledge collection	17% (17/100)
3[*]	One word sense of an NV word-pair doesn't inherit its parent category	2% (2/100)
4	The NV word-pair is not an NVEF word-pair for the sentence although it fits all confirming principles	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) has two meanings: one is <reside|住下>（緊急暫居服務(Emergent temporary residence service)）and another 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	Noun senses or verb senses that are used in Old Old Chinese	3% (3/100)
9	Word sense disambiguation failure (1) Polysemous words (2) Proper nouns identified as common words 　　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(**HwaMulan**)⇨木蘭 (**magnolia**)< FlowerGrass|花草>	27% (27/100)
10	Unknown word problem	4% (4/100)
11	Word segmentation error	2% (2/100)

[*] Types 1 through 3 no sense result from the three confirming principles of meaningful NVEF knowledge mentioned in section 3.2, respectively.

Table 8 gives examples for each type of NP-NVEF knowledge. From Tables 7, 11% of the NM-NVEF cases can be resolved by correcting the error lexicon in HowNet [Dong 1999]. The four NLP errors caused 36% NM-NVEF cases can be improved by 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]).

Table 8. Examples of eleven types of non-meaningful NVEF knowledge. The English words in parentheses are for explanatory purposes only. [] indicates nouns and <> indicates verbs.

NP type	Test Sentence	Noun / DEF	Verb / DEF
1	警方維護地方[治安]<辛勞> (Police work hard to safeguard the local 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 (The 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 conditions)	不足 (lack) 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 used to describe some one's beautiful 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 lost	存款 (bank savings) money|貨幣,$SetAside|留存	失血 (bleed or loose(only used 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,@SetAside|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] (retailer)	通路 (route) facilities|設施,route|路	零售 (retail sales) 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

4   Conclusion and Directions for Future Research

In this paper, we present an auto-generate system of NVEF knowledge (AUTO-NVEF) that fully automatically discoveries and constructs large-scale NVEF knowledge for NLP and NLU systems. AUTO-NVEF uses both human-editing knowledge (HowNet conceptual constrains) and machine-learning knowledge (word-context patterns). Experiment results show that AUTO-NVEF achieves 98.52% accuracy for news and 96.41% accuracy for specific text types. The average character number between noun and verb of NVEF knowledge is 3. Since only 2.3% of sentences in *ASBC* are N1V1-only sentences, N1V1 NVEF knowledge should be not a critical issue for NVEF-based applications. From our experimental results, word-segmentation and POS tagging both are not critical issues for our AUTO-NVEF. The critical problems, about 60% error cases, are caused by failed word-sense disambiguation (WSD) and incorrect HowNet lexicon. Therefore, conventional maximum matching word-segmentation and bi-gram like POS tagging algorithms are enough for the goal of this study. By applying AUTO-NVEF to the 2001 *UDN* corpus, we create 173,744 NVEF sense-pairs (8.8M) and 430,707 NVEF word-pairs (14.1M) on an NVEF-KR tree. Using this collection of NVEF knowledge, we incorporate 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 portions of Chinese sentences. From [Tsai *et al*. 2002] and [Wu *et al*. 2003a; Wu *et al*. 2003b], NVEF knowledge has been investigated and shown it is useful on WSD, STW, domain event extraction, domain ontology generation and text categorization.

As per our estimation, the auto-acquired NVEF knowledge from the 2001 *UDN* corpus in conjunction with the NVEF word-pair identifier [Tsai *et al*. 2002] can identify 54% and 60% of the NVEF-sentences in *ASBC* and in the 2001 *UDN* corpus, respectively. Since 94.73% (9,345/9,865) of the nouns in the most frequent 60,000 CKIP lexicon is contained in NVEF knowledge construction, the auto-generated NVEF knowledge is an acceptable large-scale NVEF knowledge for NLP/NLU systems. We found that the remaining 51.16% (5,122/10,011) of the noun-senses in HowNet are caused by two problems. One is that words with multiple noun-senses or multiple verb-senses, which are not easily resolved by WSD (say, fully-auto machine learning techniques), especially for single-character words. In our system dictionary, the maximum and average word-sense numbers of single-character words are 27 and 2.2, re-spectively. The other problem occurs from a sparse corpus. We will continue expanding our NVEF knowledge through other corpora so that we can identify more than 75% NVEF-sentences in *ASBC*. AUTO-NVEF will be extended to auto-generate other meaningful content word constructions, in particular, meaningful noun-noun, noun-adjective and

verb-adverb word-pairs. As well, we will investigate the effectiveness of NVEF knowledge in other NLP and NLU applications, such as syllable and speech understanding as well as full and shallow parsing. From [董振東 1998; Jian 2003; Dong 2004], it has been addressed that the knowledge of bilingual Verb-Noun (VN) grammatical collections, i.e. NVEF word-pairs, is a very critical issue for machine translation (MT) problem. This encourage our work on auto-generation of bilingual NVEF knowledge, especially Chinese-English, for supporting MT research fields.

5. Acknowledgements

Reference

Benson, M., E. Benson, and R. Ilson, *The BBI Combination Dictionary of English: A Guide to Word Combination*, John Benjamins, Amsterdam, Netherlands, 1986

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, C.G., K.J. Chen and L.S. Lee, "A Model for Lexical Analysis and Parsing of Chinese Sentences," *Proceedings of 1986 International Conference on Chinese Computing, Singapore*, 1986, pp.33-40

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

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

Chung, S. F., Ahrens, K., and Huang C. "ECONOMY IS A PERSON: A Chinese-English Corpora and Ontological-based Comparison Using the Conceptual Mapping Model," *In Proceedings of the 15th ROCLING Conference for the Association for Computational Linguistics and Chinese Language Processing*, National Tsing-Hwa University, Taiwan, 2003, pp.87-110

Church, K. W. and P. Hanks, "Word Association Norms, Mutual Information, and Lexicongra-

phy," *Computational Linguistics*, 1990, 16(1), pp.22-29

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

CKIP (Chinese Knowledge Information processing Group), *A study of Chinese Word Boundaries and Segmentation Standard for Information processing (in Chinese)*. Technical Report, Taiwan, Taipei, Academia Sinica, 1996.

Dang, H. T., K. Kipper and M. Palmer, "Integrating compositional semantics into a verb lexicon," *COLING-2000 Eighteenth International Conference on Computational Linguistics*, Saarbrücken, Germany, July 31 - August 4, 2000

Dong, Z. and Q. Dong, *HowNet*, http://www.keenage.com/, 1999

Dong, Z., Tutorials of HowNet, *The First International Joint Conference on Natural Language Processing (IJCNLP-04)*, 2004

Fellbaum, C., *WordNet: An Electronic Lexical Database*, MIT Press, Cambridge, MA, 1998

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

Huang, C. R., K. J. Chen, Y. Y. Yang, "Character-based Collection for Mandarin Chinese," *In ACL 2000*, 2000, pp.540-543

Huang, C. R., K. J. Chen, "Issues and Topics in Chinese Natural Language Processing," *Journal of Chinese Linguistics*, Monograph series number 9, 1996, pp.1-22

Jian, J. Y., "Extracting Verb-Noun Collections from Text," *Proceeding of ROCLING XV*, 2003, pp.295-302

Kipper K., H. T. Dang and M. Palmer, "Class-Based Construction of a Verb Lexicon," *AAAI-2000 Seventeenth National Conference on Artificial Intelligence*, Austin, TX, July 30 - August 3, 2000

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), 2000, pp.83-95

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

Lin, D., "*Using Collection Statistics in Information Extraction*," In Proc. of the Seventh Message Understanding Conference (MUC-7), 1998

Miller G., "WordNet: An On-Line Lexical Database," *International Journal of Lexicography*, 1990, 3(4)

Niles, I., and Pease, A, "Origins of the Standard Upper Merged Ontology: A Proposal for the IEEE Standard Upper Ontology," *In Working Notes of the IJCAI-2001 Workshop on the*

IEEE Standard Upper Ontology, Seattle, Washington, August 6, 2001.

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.

Smadjia, F., "Retrieving Collections from Text: Xtract," *Computational Linguistics*, 19(1), pp.143-177

Smadjia, F., K. R. McKeown, and V. Hatzivassiloglou, "Translating Collections for Bilingual Lexicons: A Statistical Approach," *Computational Linguistics*, 22(1) 1996, pp.1-38

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

Subrata D., Shuster K., and Wu, C., "Ontologies for Agent-Based Information Retrieval and Sequence Mining," *In Proceedings of the Workshop on Ontologies in Agent Systems (OAS02)*, held at the 1st International Joint Conference on Autonomous Agents and Multi-Agent Systems Bologna, Italy, July, 2002, pp.15-19

Sun, J., J. Gao, L. Zhang, M. Zhou and C. Huang, "Chinese Named Entity Identification Using Class-based Language Model," *In the Proceedings of 19th COLING 2002*, Taipei, 2000, 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, 2002, pp.1016-1022

Tsai, J. L, C. L. Sung and W. L. Hsu, "Chinese Word Auto-Confirmation Agent," *In Proceedings of the 15th ROCLING Conference for the Association for Computational Linguistics and Chinese Language Processing*, National Tsing-Hwa University, Taiwan, 2003, pp.175-192

Wu, S. H., T. H. Tsai, and W. L. Hsu, "Text Categorization Using Automatically Acquired Domain Ontology," *In proceedings of the Sixth International Workshop on Information Retrieval with Asian Languages (IRAL-03)*, Sapporo, Japan, 2003, pp.138-145

Wu, S. H., T. H. Tsai, and W. L. Hsu, "Domain Event Extraction and Representation with Domain Ontology," *In proceedings of the IJCAI-03 Workshop on Information Integration on the Web*, Acapulco, Mexico, 2003, pp.33-38

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

朱曉亞，*現代漢語句模研究(Studies on Semantic Structure Patterns of Sentence in Modren Chinese)*，北京大學出版社，2001

胡裕樹，范曉，*動詞研究*，河南大學出版社，1995

董振東，語義關係的表達和知識系統的建造，*語言文字應用*，1998，第 3 期，頁 76-82

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

陳昌來，*現代漢語動詞的句法語義屬性研究(XIANDAI HANYU DONGCI DE JUFA YUYI SHUXING YANJIU)*，學林出版社，2002

劉順，*現代漢語名詞的多視角研究(XIANDAI HANYU DONGCI DE JUFA YUYI SHUXING YANJIU)*，學林出版社，2003

Appendix A. Sample Table of Main Noun Features and Noun-Sense Classes

Main noun features	Noun-sense classes
bacteria|微生物	微生物(bacteria)
AnimalHuman|動物	動物類(animal)
human|人	人物類(human)
plant|植物	植物類(plant)
artifact|人工物	人工物(artifact)
natural|天然物	天然物(natural)
fact|事情	事件類(event)
mental|精神	精神類(mental)
phenomena|現象	現象類(phenomena)
shape|物形	物形類(shape)
InstitutePlace|場所	地點類(place)
location|位置	位置類(location)
attribute|屬性	抽象類(abstract)
quantity|數量	數量類(quantity)

Appendix B. Example Mappings of FPOS and NV Word-Pairs

FPOS	NV word-pairs	Example, [] indicates nouns and <> indicates verbs
$N_1V_2ADJ_3N_4$	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 NVEF accepting Conditions

Noun-sense class	Verb DEF	Example, [] indicates nouns and <> indicates verbs
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

Appendix D. User Interface for Confirming generated NVEF Knowledge