Building the Chinese Open Wordnet (COW): Starting from Core Synsets

Shan Wang, Francis Bond

Division of Linguistics and Multilingual Studies,
Nanyang Technological University,
14 Nanyang Drive, Singapore 637332

wangshanstar@gmail.com, bond@ieee.org

Abstract

Princeton WordNet (PWN) is one of the most influential resources for semantic descriptions, and is extensively used in natural language processing. Based on PWN, three Chinese wordnets have been developed: Sinica Bilingual Ontological Wordnet (BOW), Southeast University WordNet (SEW), and Taiwan University WordNet (CWN). We used SEW to sense-tag a corpus, but found some issues with coverage and precision. We decided to make a new Chinese wordnet based on SEW to increase the coverage and accuracy. In addition, a small scale Chinese wordnet was constructed from open multilingual wordnet (OMW) using data from Wiktionary (WIKT). We then merged SEW and WIKT. Starting from core synsets, we formulated guidelines for the new Chinese Open Wordnet (COW). We compared the five Chinese wordnets, which shows that COW is currently the best, but it still has room for further improvement, especially with polysemous words. It is clear that building an accurate semantic resource for a language is not an easy task, but through consistent efforts, we will be able to achieve it. COW is released under the same license as the PWN, an open license that freely allows use, adaptation and redistribution.

1 Introduction

Semantic descriptions of languages are useful for a variety of tasks. One of the most influential such resources is the Princeton WordNet (PWN), an English lexical database created at the Cognitive Science Laboratory of Princeton University (Fellbaum, 1998; George A Miller, 1995; George A. Miller, Beckwith, Fellbaum, Gross, & Miller, 1990). It is widely used in natural language processing tasks, such as word sense disambiguation, information retrieval and text classification. PWN has greatly improved the performance of these tasks. Based on PWN, three Chinese wordnets have been developed. Sinica Bilingual Ontological Wordnet (BOW) was created through a bootstrapping method (Huang, Chang, & Lee, 2004; Huang, Tseng, Tsai, & Murphy, 2003). Southeast University Chinese WordNet (SEW) was automatically constructed by implementing three approaches, including Minimum Distance, Intersection and Words Co-occurrence (Xu, Gao, Pan, Qu, & Huang, 2008); Taiwan University and Academia Sinica also developed a Chinese WordNet (CWN)(Huang *et al* 2010). We used SEW to sense-tag NTU corpus data (Bond, Wang, Gao, Mok, & Tan, 2013; Tan & Bond, 2012). However, its mistakes and its coverage hinder the progress of the sense-tagged corpus. Moreover, the open multilingual wordnet project (OMW)[1] created wordnet data for many languages, including Chinese (Bond & Foster, 2013). Based on OMW, we created a small scale Chinese wordnet from Wiktionary (WIKT).

All of these wordnets have some flaws and, when we started our project, none of them were available under an open license. A high-quality and freely available wordnet would be an important resource for the community. Therefore, we have started work on yet another Chinese wordnet in Nanyang Technological University (NTU COW), aiming to produce one with even better accuracy and coverage. Core synsets[2] are the most common ones ranked according to word frequency in British National Corpus (Fellbaum & Vossen, 2007). There are 4,960 synsets after mapping to WordNet 3.0. These synsets are more salient than others, so we began with them.

In this paper we compared all the five wordnets (COW, BOW, SEW, WIKT, and CWN), and showed their strengths and weaknesses.

The following sections are organized as follows.

[1] http://www.casta-net.jp/~kuribayashi/multi/

[2] http://wordnet.cs.princeton.edu/downloads.html

International Joint Conference on Natural Language Processing, pages 10–18,
Nagoya, Japan, 14-18 October 2013.

Section 2 elaborates on the four Chinese wordnets built based on PWN. Section 3 introduces the guidelines in building COW. Section 4 compares the core synsets of different wordnets. Finally the conclusion and future work are stated in Section 5.

2 Related Research

PWN was developed from 1985 under the direction of George A. Miller. It groups nouns, verbs, adjective and adverbs into synonyms (synsets), most of which are linked to other synsets through a number of semantic relations. For example, nouns have these relations: hypernym, hyponym, holonym, meronym, and coordinate term (Fellbaum, 1998; George A Miller, 1995; George A. Miller et al., 1990). PWN has been a very important resource in computer science, psychology, and language studies. Hence many languages followed up and multilingual wordnets were either under construction or have been built. PWN is the mother of all wordnets (Fellbaum, 1998). Under this trend, in the Chinese community, three wordnets were built: SEW, BOW, and CWN. SEW is in simplified Chinese, while BOW and CWN are in traditional Chinese.

SEW:[3] Xu et al. (2008) investigated various automatic approaches to translate the English WordNet 3.0 to Chinese WordNet. They are Minimum Distance (MDA), Intersection (IA) and Words Co-occurrence (WCA). MDA computes the Levenshtein Distance between glosses of English synsets and the definition in American Heritage Dictionary (Chinese & English edition). IA chooses the intersection of the translated words. WCA put an English word and a Chinese word as a group to get the co-occurrence results from Google. IA has the highest precision, but the lowest recall. WCA has highest recall but lowest recall. Considering the pros and cons of each approach, they then integrated them into an integrated one called MIWA. They first chose IA to process the whole English WordNet then MDA to deal with the remaining synsets of WordNet; finally adopt WCA for the rest. Following this order, MIWA got a high translation precision and increased the number of synsets that can be translated. SEW is free for research, but cannot be redistributed.

BOW:[4] It was bootstrapped from the English-Chinese Translation Equivalents Database (ECTED), based on WordNet 1.6(Huang et al., 2003; Huang, Tseng, & Tsai, 2002). ECTED was manually made by the Chinese Knowledge and Information Processing group (CKIP), Academia Sinica. First, all Chinese translations of an English lemma from WordNet 1.6 were extracted from online bilingual resources. They are checked by a team of translators who select the three most appropriate translation equivalents where possible (Huang et al., 2004). They tested the 210 most frequent Chinese lexical lemmas in Sinica Corpus. They first mapped them to ECTED to find out their corresponding English synsets and then by assuming the WordNet semantic relations hold true for Chinese, they automatically linked the semantic relations for Chinese. They further evaluated the semantic relations in Chinese, which showed that automatically assigned relation in Chinese has high probability once the translation is equivalent (Huang et al., 2003). BOW is only available for online lookup.

CWN:[5] BOW has many entries that are not truly lexicalized in Chinese. To solve this issue, Taiwan University constructed a Chinese wordnet with the aim of making only entries for Chinese words (Huang et al., 2010). CWN was recently released under the same license as wordnet.

Besides the above three Chinese wordnets, we looked at data from Bond and Foster (2013) who extracted lemmas for over a hundred languages by linking the English Wiktionary to OMW (WIKT). By linking through multiple translations, they were able to get a high precision for commonly occurring words. For Chinese, they found translations for 12,130 synsets giving 19,079 senses covering 49% of the core synsets.

We did some cleaning up and mapped the above four wordnets into WordNet 3.0. The size of each one is depicted in Table 1. SEW has the most entries, followed by BOW. SEW, BOW and WIKT have nouns as the largest category, while CWN has verbs as the largest category.

3 Build the Chinese Open Wordnet

We have been using SEW to sense-tag the Chinese part of the NTU Multi-Lingual Corpus

[3] http://www.aturstudio.com/wordnet/windex.php

[4] http://bow.sinica.edu.tw/wn/

[5] http://lope.linguistics.ntu.edu.tw/cwn/query/

which has 6,300 sentences from texts of different genres: (i) two stories: *The Adventure of the Dancing Men*, and *The Adventure of the Speckled Band*; (ii) an essay: The Cathedral and the Bazaar; (iii) news: Mainichi News; and (iv) tourism: Your Singapore (Tan & Bond, 2012). However, as SEW is automatically constructed, it was found that there are many mistakes and some words are not included.

POS	SEW		BOW		CWN		WIKT	
	No.	Percent (%)	No.	Percent (%)	No.	Percent(%)	No.	Percent(%)
noun	100,064	63.7	91,795	62.3	2822	32.6	14,976	78.5
verb	22,687	14.4	20,472	13.9	3676	42.5	2,128	11.2
adjective	28,510	18.1	29,404	20.0	1408	16.3	1,566	8.2
adverb	5,851	3.7	5,674	3.9	747	8.6	409	2.1
Total	157,112	100.0	147,345	100.0	8,653	100.0	19,079	100.0

Table 1. Size of SEW, BOW, CWN, and WIKT

In order to ensure coverage of frequently occurring concepts, we decided to concentrate on the core synsets first, following the example of the Japanese wordnet (Isahara, Bond, Uchimoto, Utiyama, & Kanzaki, 2008). The core synsets of PWN are the most frequent nouns, verbs, and adjectives in British National Corpus (BNC)[6] (Boyd-Graber, Fellbaum, Osherson, & Schapire, 2006). There are 4,960 synsets after mapping them to WordNet 3.0. Nouns are the largest category making up to 66.1%. Verbs account for 20.1% and adjectives only take up 13.8%. There is no adverb in the core synsets.

The construction procedure of COW comprises of three phases: (i) extract data from Wiktionary and then merge WIKT and SEW, (ii) manually check all translations by referring to bilingual dictionaries and add more entries, (iii) check the semantic relations. The following section introduces the phases.

COW is released under the same license as the PWN, an open license that freely allows use, adaptation and redistribution. Because SEW, WIKT and the corpus we are annotating are in simplified Chinese, COW is also made in simplified Chinese.

3.1 Merge SEW and WIKT

We were able to obtain a research license for SEW. WIKT data is under the same license as Wiktionary (CC BY SA[7]) and so can be freely used. We merged the two sets and extracted only the core synsets, which gave us a total of 12,434 Chinese translations for the 4,960 core synsets.

3.2 Manual Correction of Chinese Translations

During the process of manual efforts in building a better Chinese wordnet, we drew up some guidelines. First, Chinese translations must convey the same meaning and POS as the English synset. If there is a mismatch in senses, transitivity and POS (not including cases that need to add 的 *de* / 地 *de*), delete it. Second, use simplified and correct orthography. If the Chinese translations must add 的 *de* /地 *de* to express the same POS as English, add it. The second guideline is referred to as amendments. Third, add new translations through looking up authoritative bilingual dictionaries. The following section describes the three actions taken (delete, amend, and add) by using the three guidelines.

3.2.1 Delete a Wrong Translation

A translation will be deleted if it is in one of the three cases: (i) wrong meaning; (ii) wrong transitivity; (iii) wrong POS.

[6] http://www.natcorp.ox.ac.uk/

[7] Creative Commons: Attribution-ShareAlike, http://creativecommons.org/licenses/by-sa/3.0/

(i) Wrong Meaning

If a Chinese translation does not reflect the meaning of an English synset, delete it. For instance, *election* is a polysemous word, which has four senses in PWN:

- S1: (n) **election** (a vote to select the winner of a position or political office) *"the results of the election will be announced tonight"*
- S2: (n) **election** (the act of selecting someone or something; the exercise of deliberate choice) *"her election of medicine as a profession"*
- S3: (n) **election** (the status or fact of being elected) *"they celebrated his election"*
- S4: (n) **election** (the predestination of some individuals as objects of divine mercy (especially as conceived by Calvinists))

The synset 00181781-n is the first sense of “election” (S1) in WordNet. The Chinese WordNet provides two translations: 当选 *dāngxuǎn* ‘election’ and 选举 *xuǎnjǔ* ‘election’. It is clear that 当选 *dāngxuǎn* ‘election’ is the third sense of “election”, so it should be deleted.

(ii) Wrong Transitivity

Verbs usually have either transitive or intransitive use. In synset 00250181-v, “mature; maturate; grow” are intransitive verbs, so the Chinese translation 使成熟 *shǐ chéngshú* ‘make mature’ is wrong and is thus deleted.

00250181-v *mature; maturate; grow* “develop and reach maturity; undergo maturation”: *He matured fast; The child grew fast*

(iii) Wrong POS

When the POS of an English synset has a Chinese translation that has the same POS, then the Chinese translation with a different POS should be deleted. For example, 00250181-v is a verbal synset, but 壮年的 *zhuàngnián de* ‘the prime of life’s’ and 成熟的 *chéngshú de* ‘mature’ are not verbs, so they are deleted.

3.2.2 Amend a Chinese Translation

A translation will be amended if it is in one of the three cases: (i) written in traditional characters; (ii) wrong characters; (iii) need 的 *de* /地 *de* to match the English POS.

(i) Written in Traditional Characters

When a Chinese translation is written in traditional Chinese, amend it to be simplified Chinese. The synset 02576460-n is translated as 鰺屬 *shēn shǔ* ‘caranx’, we change it to be 鲹属 *shēn shǔ* ‘caranx’.

02576460-n *Caranx; genus_Caranx* “type genus of the Carangidae”

(ii) Wrong Characters

When a Chinese translation has a typo, revise it to the correct one. The synset 00198451-n is translated as 晋什 *jìnshén*, which should have been 晋升 *jìnshēng* ‘promotion’.

00198451-n *promotion* “act of raising in rank or position”

(iii) Need 的 *de* /地 *de* to match the English POS

The synset 01089369-a is an adjectival, but the translation 兼职 *jiānzhí* ‘part time’ is a verb/noun, so we add 的 *de* to it (1.3).

01089369-a *part-time; part time* “involving less than the standard or customary time for an activity”: *part-time employees; a part-time job*

3.2.3 Add Chinese Translations

To improve the coverage and accuracy of COW, we make reference not only to many authoritative bilingual dictionaries, such as The American Heritage Dictionary for Learners of English (Zhao, 2006), The 21st Century Unabridged English-Chinese Dictionary (Li, 2002), Collins COBUILD Advanced Learner's English-Chinese Dictionary (Ke, 2011), Oxford Advanced Learner's English-Chinese Dictionary (7th Edition) (Wang, Zhao, & Zou, 2009), Longman Dictionary of Contemporary English (English-Chinese) (Zhu, 1998), etc., but also online bilingual dictionaries, such as iciba[8], youdao[9], lingoes[10], dreye[11] and bing[12].

For example, the English synset *00203866-v* can be translated as 变坏 *biàn huài* ‘decline’ and 恶化 *èhuà* ‘worsen’, which are not available in the current wordnet, so we added them to COW.

00203866-v *worsen; decline* “grow worse”: *Conditions in the slum worsened*

3.3 Check Semantic Relations

PWN groups nouns, verbs, adjectived and adverbs

[8] http://www.iciba.com/
[9] http://dict.youdao.com/
[10] http://www.lingoes.cn/
[11] http://www.dreye.com.cn/
[12] http://cn.bing.com/dict/

into synonyms (synsets), most of which are linked to other synsets through a number of semantic relations. Huang et al. (2003) tested 210 Chinese lemmas and their semantic relations links. The results show that lexical semantic-relation translations are highly precise when they are logically inferable. We randomly checked some of the relations in COW, which shows that this statement also holds for the new Chinese wordnet we are building.

3.4 Results of the COW Core Synsets

Through merging SEW and WIKT, we got 12,434 Chinese translations. Based on the guidelines described above, the revisions we made are outlined in Table 2.

Wrong Entries	Deletion	1,706
	Amendment	134
Missing Entries	Addition	2,640
Total		4,480

Table 2. Revision of the wordnet

Table 2 shows that there are 1,840 wrong entries (15%) of which we deleted 1,706 translations and amended 134. Furthermore, we added 2,640 new entries (about 21%).

The wrong entries are further checked according to POS as shown in Table 3. The results indicate that verbal synsets have a higher error rate than nouns and adjectives. This is because verbs tend to be more complex than words in other grammatical categories. This also reminds us to pay more attention to verbs in building the new wordnet.

Synset POS	Wrong Entries		All Entries		Error Rate (Wrong/All)
	No.	Percent(%)	No.	Percent(%)	Percent(%)
Noun	1,164	63.3	7,823	62.9	14.9
Verb	547	29.7	3,087	24.8	17.7
Adjective	129	7.0	1,524	12.3	8.5
Total	1,840	100.0	12,434	100.0	14.8

Table 3. Error rate of entries by POS

4 Compare Core Synsets of Five Chinese Wordnets

Many efforts have been devoted to the construction of Chinese wordnets. To get a general idea of the quality of each wordnet, we randomly chose 200 synsets from the core synsets of the five Chinese wordnets and manually made gold standard for Chinese entries. During this process, we noticed that due to language difference, it is hard to make a decision for some cases. In order to better compare the synset lemmas, we created both a strict gold standard and a loose gold standard.

4.1 Creating Gold Standards

This section discusses the gold standard from word meaning, POS and word relation.

4.1.1 Word Meaning

Leech (1974) recognized seven types of meaning: conceptual meaning, connotative meaning, social meaning, affective meaning, reflected meaning, collocative meaning and thematic meaning. Fu (1985) divided word meaning into conceptual meaning and affiliated meaning. The latter is composed of affective color, genre color and image color. Liu (1990) divided word meaning into conceptual meaning and color meaning. The latter is further divided into affective color, attitude color, evaluation color, image color, genre color, style color, (literary or artistic) style color and tone color. Ge (2006) divided word meaning into conceptual meaning, color meaning and grammatical meaning.

Following these studies, the following section divides word meaning into conceptual meaning and affiliated meaning. Words with similar conceptual meaning may differ in the *meaning severity* and *the scope of meaning usage*. Regarding affiliated meaning, words may differ in affection, genre and time of usage.

4.1.1.1 Conceptual Meaning

Some English synset have exact equivalents in Chinese. For example, the following synset 02692232-n has a precise Chinese equivalent 机场 *jīchǎng* ‘airport’.

02692232-n *airport; airdrome; aerodrome; drome* “an airfield equipped with control tower and hangars as well as accommodations for passengers and cargo”

However, in many cases, words of two languages may have similar basic conceptual meaning, but the meanings differ in severity and usage scope.

(i) Meaning Severity

Regarding the synset 00618057-v, 出错 *chūcuò* and 犯错 *fàncuò* are equivalent translation. In contrast, 失足 *shīzú* 'make a serious mistake' is much stronger and should be in a separate synset.

00618057-v *stumble; slip up; trip up* "make an error": *She slipped up and revealed the name*

(ii) Usage Scope of Meaning

For the synset 00760916-a, no Chinese lemma has as wide usage as "direct". Thus all the Chinese translations, such as 直达 *zhídá* 'directly arriving' and 直接 *zhíjiē* 'direct' have a narrower usage scope.

00760916-a *direct* "direct in spatial dimensions; proceeding without deviation or interruption; straight and short": *a direct route; a direct flight; a direct hit*

4.1.1.2 Affiliated Meaning

With respect to affiliated meaning, words may differ in affection, genre and time of usage.

(i) Affection

The synset 09179776-n refers to "positive" influence, so 激励 *jīlì* 'incentive' is a good entry. The word 刺激 *cìjī* 'stimulus' is not necessarily "positive".

09179776-n *incentive; inducement; motivator* "a positive motivational influence"

(ii) Genre

In the synset 09823502-n, the translations 妗 *jìn* 'aunt' and 妗母 *jìnmǔ* 'aunt' are Chinese dialects .

09823502-n *aunt; auntie; aunty* "the sister of your father or mother; the wife of your uncle"

(iii) Time: modern vs. ancient

In the synset 10582154-n, the translations 侍从 *shìcóng* 'servant', 仆人 *púrén* 'servant', 侍者 *shìzhě* 'servant' are used in ancient or modern China, rather than contemporary China. The word now used is 保姆 *bǎomǔ* 'servant' .

10582154-n *servant; retainer* "a person working in the service of another (especially in the household)"

4.1.2 Part of Speech (POS)

The Chinese entries should have the same POS as the English synset. In the synset 00760916-a, the translated word 径直 *jìngzhí* 'directly' is an adverb, which does not fit this synset.

00760916-a *direct* "direct in spatial dimensions; proceeding without deviation or interruption; straight and short": *a direct route; a direct flight; a direct hit*

4.1.3 Word Relations

One main challenge concerning word relations is hyponyms and hypernyms. In making our new wordnet and creating the loose gold standard, we treat the close hyponyms and close hypernyms as right, and the not so close ones as wrong. In the strict gold standard, we treat all of them as wrong.

(i) Close Hyponym

The synset 06873139-n can refer to either the highest female voice or the voice of a boy before puberty. There is no single word with the two meanings in Chinese. The translation 女高音 *nǚ gāoyīn* 'the highest female voice' is a close hyponym of this synset. For cases like this, we would create two synsets for Chinese in the future.

06873139-n *soprano* "the highest female voice; the voice of a boy before puberty"

(ii) Not Close Hyponym

The synset 10401829-n has good equivalences 参与者 *cānyùzhě* 'participant' and 参加者 *cānjiāzhě* 'participant' in Chinese. The translation 与会者 *yùhuìzhě* 'people attending a conference' refers to the people attending a conference, which is not a close hyponym.

10401829-n *participant* "someone who takes part in an activity"

(iii) Close Hypernym

The synset 02267060-v has good equivalents 花 *huā* 'spend' and 花费 *huāfèi* 'spend'. It is also translated as 使 *shǐ* 'use' and 用 *yòng* 'use', which are close hypernyms. It is possible that the two hypernyms are so general that their most typical synset does not have the meaning of spending money.

02267060-v *spend; expend; drop* "pay out": *spend money*

(iv) Not Close Hypernym

The synset 02075049-v has good equivalents such as 逃走 *táozǒu* 'scat' and 逃跑 *táopǎo* 'scat'. Meanwhile, it is translated to 跑 *pǎo* 'run' and 奔 *bēn* 'rush', which are not so close hypernyms. It is certain that to flee is to run, but the two hypernyms should have their own more suitable synsets.

02075049-v *scat; run; scarper; turn_tail; lam; run_away; hightail_it; bunk; head_for_the_hills;*

take_to_the_woods; escape; fly_the_coop; break_away "flee; take to one's heels; cut and run": *If you see this man, run!; The burglars escaped before the police showed up*

4.1.4 Grammatical Status

Lexicalization is a process in which something becomes lexical (Lehmann, 2002). Due to historical and cultural reasons, different language lexicalizes different language elements. For example, there is no lexicalized word for the synset 02991555-n in Chinese. In Chinese, you must use a phrase or definition to mean what this synset expresses.

02991555-n *cell; cubicle* "small room in which a monk or nun lives"

Considering the differences among languages, we created two gold standards for 200 randomly chosen synsets: the strict gold standard and the loose gold standard. The former aims to find the best translation for a synset; while the latter finds the correct translation. The former has some disadvantages: it makes many Chinese words not have a corresponding synset in PWN; further, it makes many English synsets have no Chinese entry. The latter solves the problems, but it is not as accurate as the former. Table 4 summarizes the action taken for creating loose and strict gold standards, as well as showing our standard in making the new wordnet. The gold standard data was created by the authors in consultation with each other. Ideally it would be better if we got multiple annotators to provide inter-annotator agreement, but the current results are derived through discussion and making reference to many bilingual dictionaries and we have come to an agreement on them.

Standard		Chinese	Loose	Strict	Making New Wordnet
Meaning	Conceptual Meaning	different from English synset	wrong	wrong	wrong
		exact equivalent	right	right	keep
		Severity	right	wrong	keep
		Usage scope	right	wrong	keep
	Affiliated Meaning	Affection: different	right	wrong	keep
		Genre: dialect	right	wrong	keep
		Time: non-contemporary	not include	wrong	keep
POS		same POS as English	right	right	keep
		no same POS as English	right	wrong	wrong
Word Relation		close hyponym/hypernym	right	wrong	keep
		not close hyponym/hypernym	wrong	wrong	wrong
Grammatical Status		word	right	right	keep
		phrase	not include	not include	keep
		morpheme	not include	not include	keep
		definition	not include	not include	keep
Orthography		wrong character	wrong	wrong	amend

Table 4. Summary of standard

4.2 Results, Discussion and Future Work

We did some cleaning up before doing evaluation, including strip off 的 *de* /地 *de* at the end of a lemma, and the contents within parentheses. We also transferred the traditional characters in BOW and CWN to simplified characters. Through applying the standards illustrated in Table 1, we evaluated the dataset through counting the precision, recall and F-score.

$$\text{Precision} = \frac{\text{No.of correct lemmas in each core synset}}{\text{No.of all lemmas in each core synsets}}$$

$$\text{Recall} = \frac{\text{No.of correct lemmas in each core synset}}{\text{No.of correct lemmas in all core synsets}}$$

$$\text{F-score} = 2*\frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

The results of using the loose and strict gold standards are indicated in Table 5 and Table 6 respectively. All wordnets were tested on the same samples described above.

Wordnet	COW	BOW	SEW	WIKT	CWN
precision	0.86	0.80	0.75	0.92	0.56
recall	0.77	0.48	0.45	0.32	0.08
F-score	0.81	0.60	0.56	0.47	0.14

Table 5. Loose gold standard

Wordnet	COW	BOW	SEW	WIKT	CWN
precision	0.81	0.76	0.70	0.88	0.46
recall	0.80	0.50	0.46	0.33	0.07
F-score	0.81	0.60	0.55	0.48	0.13

Table 6. Strict gold standard

The results of the two standards show roughly the same F-score: the strict/loose distinction does not have large effect. This is because there were few entries where the loose and strict gold standards actually differ. By using the strict gold standard, the recall of each wordnet increased except CWN. Meanwhile, the precision of each wordnet decreased.

COW was built using the results of both SEW and WIKT along with a lot of extra checking. It is therefore not surprising that it got the best precision and recall. Exploiting data from multiple existing wordnets makes a better resource. BOW ranked second according to the evaluation. It was bootstrapped from a translation equivalence database. Though this database was manually checked, it cannot guarantee that they will give an accurate wordnet. SEW and WIKT were automatically constructed and thus have low F-score, but WIKT has high precision. This is because it was created using 20 languages to disambiguate the meaning instead of only looking at English and Chinese. CWN turned out to have the lowest score. This is because the editors are mainly focusing on implementing new theories of complex semantic types and not aiming for high coverage.

Among all the five wordnets we compared, COW is the best according to the evaluation. However, even though both it and BOW were carefully checked by linguists, there are still some mistakes, which show the difficulty in creating a wordnet. The errors mainly come from the polysemous words, which may have been assigned to another synset. One reason leading to such errors comes from the fact that core synsets alone do not show all the senses of a lemma. If a lemma is divided into different senses especially when they are fine-grained and only one of the senses is presented to the editors, it is hard to decide which is the best entry for another language. What we have done with the core synsets is a trial to find the problems and test our method. It is definitely not enough to go through all the data once, and thus we will further revise all the wrong lemmas. By taking the core synset as the starting point of our large-scale project on constructing COW, we not only got more insight into language disparities between English and Chinese, but also become clearer about what rules to take in constructing wordnets, which will in turn benefit the construction of other high-quality wordnets.

In further efforts we are validating the entries by sense tagging parallel corpora (Bond et al, 2013): this allows us to see the words in use and compare them to wordnets in different languages. Monolingually, it allows us to measure the distribution of word senses. With the construction of a high-accuracy, high-coverage Chinese wordnet, it will not only promote the development of Chinese Information Processing, but also improve the combined multilingual wordnet.

We would also like to investigate making wordnet in traditional characters as default and automatically converting to simplified (it is lossy in the other direction).

5 Conclusions

This paper introduced our on-going work of building a new Chinese Open wordnet: NTU COW. Due to language divergence, we met many theoretical and practical issues. Starting from the core synsets, we formulated our guidelines and become clearer about how to make a better wordnet. Through comparing the core synsets of five wordnets, the results show that our new wordnet is the current best. Although we carefully checked the core synsets, however, we still spotted some errors which mainly come from selecting the suitable sense of polysemous words. This leaves us space for more improvement and gives us a lesson about how to make the remaining parts much better. The wordnet is open source, so the data can be used by anyone at all, including the other wordnet projects.

Acknowledgments

This research was supported by the MOE Tier 1 grant *Shifted in Translation—An Empirical Study of Meaning Change across Languages* (2012-T1-001-135) and the NTU HASS Incentive Scheme *Equivalent But Different: How Languages Represent Meaning In Different Ways*.

References

Bond, Francis, & Foster, Ryan. (2013). Linking and Extending an Open Multilingual Wordnet *Proceedings of The 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)* (pp. 1352-1362). Sofia, Bulgaria.

Bond, Francis, Wang, Shan, Gao, Huini, Mok, Shuwen, & Tan, Yiwen. (2013). Developing Parallel Sense-tagged Corpora with Wordnets *Proceedings of the 7th Linguistic Annotation Workshop & Interoperability with Discourse, Workshop of The 51st Annual Meeting of the Association for Computational Linguistics (ACL-51)* (pp. 149-158). Sofia, Bulgaria.

Boyd-Graber, Jordan, Fellbaum, Christiane, Osherson, Daniel, & Schapire, Robert. (2006). *Adding dense, weighted, connections to WordNet.* Paper presented at the Proceedings of the Third International WordNet Conference.

Fellbaum, Christiane. (1998). *Wordnet: An Electronic Lexical Database*. MA: MIT Press.

Fellbaum, Christiane, & Vossen, Piek. (2007). Connecting the Universal to the Specific: Towards the Global Grid. In Toru Ishida, Susan R. Fussell & Piek T. J. M. Vossen (Eds.), *Intercultural Collaboration: First International Workshop on Intercultural Collaboration (IWIC-1)* (Vol. 4568, pp. 2-16). Berlin-Heidelberg: Springer.

Fu, Huaiqing. (1985). *Modern Chinese Lexicon (现代汉语词汇)*: Peking University Press.

Ge, Benyi. (2006). *Research on Chinese Lexicon (汉语词汇研究)*. Beijing: Foreign Language Teaching and Research Press.

Huang, Chu-Ren, Chang, Ru-Yng, & Lee, Shiang-Bin. (2004). Sinica BOW (Bilingual Ontological Wordnet): Integration of Bilingual WordNet and SUMO *Proceedings of the 4th International Conference on Language Resources and Evaluation* (pp. 1553-1556).

Huang, Chu-Ren, Hsieh, Shu-Kai, Hong, Jia-Fei, Chen, Yun-Zhu, Su, I-Li, Chen, Yong-Xiang, & Huang, Sheng-Wei. (2010). Chinese WordNet: Design and Implementation of a Cross-Lingual Knowledge Processing Infrastructure. *Journal of Chinese Information Processing, 24*(2), 14-23.

Huang, Chu-Ren, Tseng, Elanna I. J., Tsai, Dylan B. S., & Murphy, Brian. (2003). Cross-lingual Portability of Semantic Relations: Bootstrapping Chinese WordNet with English WordNet Relations. *Language and Linguistics, 4*(3), 509-532.

Huang, Chu-Ren, Tseng, Elanna I.J., & Tsai, Dylan B.S. (2002). *Translating Lexical Semantic Relations: The First Step Towards Multilingual Wordnets.* Paper presented at the Proceedings of the Workshop on Semanet: Building and Using Semantic Networks: COLING 2002 Post-conference Workshops, Taipei.

Isahara, Hitoshi, Bond, Francis, Uchimoto, Kiyotaka, Utiyama, Masao, & Kanzaki, Kyoko. (2008). Development of the Japanese WordNet *Proceedings of The Sixth International Conference on Language Resources and Evaluation (LREC-6)*. Marrakech.

Ke, Ke'er. (Ed.) (2011) Collins COBUILD Advanced Learner's English-Chinese Dictionary. Beijing: Foreign Language Teaching and Research Press & Harper Collins Publishers Ltd.

Leech, Geoffrey N. (1974). *Semantics*. London: Penguin.

Lehmann, Christian. (2002). Thoughts on Grammaticalization.

Li, Huaju. (Ed.) (2002) The 21st Century Unabridged English-Chinese Dictionary. Beijing: China Renmin University Press Co., LTD.

Liu, Shuxin. (1990). *Chinese Descriptive Lexicology (汉语描写词汇学)*. The Commercial Press.

Miller, George A. (1995). WordNet: a lexical database for English. *Communications of the ACM, 38*(11), 39-41.

Miller, George A., Beckwith, Richard, Fellbaum, Christiane, Gross, Derek, & Miller, Katherine J. (1990). Introduction to wordnet: An online lexical database. *International journal of lexicography, 3*(4), 235-244.

Tan, Liling, & Bond, Francis. (2012). Building and annotating the linguistically Diverse NTU-MC (NTU-multilingual corpus). *International Journal of Asian Language Processing, 22*(4), 161–174

Wang, Yuzhang, Zhao, Cuilian, & Zou, Xiaoling. (Eds.). (2009) Oxford Advanced Learner's English-Chinese Dictionary (7th Edition). Beijing: The Commercial Press & Oxford University Press.

Xu, Renjie, Gao, Zhiqiang, Pan, Yingji, Qu, Yuzhong, & Huang, Zhisheng. (2008). An integrated approach for automatic construction of bilingual Chinese-English WordNet. In John Domingue & Chutiporn Anutariya (Eds.), *The Semantic Web: 3rd Asian Semantic Web Conference* (Vol. 5367, pp. 302-314): Springer.

Zhao, Cuilian. (Ed.) (2006) The American Heritage Dictionary for Learners of English. Beijing: Foreign Language Teaching and Research Press & Houghton Mifflin Company.

Zhu, Yuan. (Ed.) (1998) Longman Dictionary of Contemporary English (English-Chinese). Beijing: The Commerical Press & Addison Wesley Longman China Limited.