Issues in building English-Chinese parallel corpora with WordNets

Francis Bond and **Shan Wang**

Linguistics and Multilingual Studies,
Nanyang Technological University
bond@ieee.org,wangshanstar@gmail.com

Abstract

We discuss some of the issues in producing sense-tagged parallel corpora: including pre-processing, adding new entries and linking. We have preliminary results for three genres: stories, essays and tourism web pages, in both Chinese and English.

1 Introduction

Since the first release of the Princeton WordNet (PWN) (Fellbaum, 1998) there has been a great increase in the size and number of wordnets created (Bond and Paik, 2012). Further, there has been an empirical revolution in natural language processing (Vanderwende and Menezes, 2005), with machine learning based on annotated corpora dominating the field. Given this, we would expect to see a flowering of sense annotated corpora. However, they are still relatively rare and small in size compared to part-of-speech and tree banked corpora (Petrolito and Bond, 2014).

In this paper we describe ongoing work to sense annotate data in two languages (English and Chinese), using texts provided by the *Nanyang Technological University Multilingual Corpus* (NTU-MC: Tan and Bond, 2012). We discuss some of the problems involved with pre-processing (Section 3), monolingual sense tagging (Section 4) and multi-lingually linking the data (Section 5). We then discuss some ideas to improve the annotation process (Section 6) and conclude.

2 Related Research

Sense-tagged parallel corpora are an important resource for NLP, contrastive linguistics and bilingual lexicography. However, there are few multilingual sense tagged corpora. One notable exception is the MultiSemCor (Pianta et al., 2002). Taking the English SemCor (Landes et al., 1998) as a source, first Italian, then Romanian and Japanese translations have been made. The leading project was the Italian SemCor with 268,905 Italian tokens and 258,499 English tokens (Pianta et al., 2002). This was followed by the Romanian SemCor with 175,603 tokens in Romanian matched with 178,499 English tokens (Lupu et al., 2005). Finally, the Japanese SemCor has senses projected across from English. Of the 150,555 content words, 58,265 are sense tagged either as monosemous words or by projecting from the English annotation (Bond et al., 2012).

Some universities have devoted efforts to construct Chinese-English parallel corpora, such as Peking University, Tsinghua University and Chinese Academy of Sciences (Chang et al., 2003; Chang, 2004), Xiamen University (Chen et al., 2005, 2006), Beijing Foreign Studies University (Wang, 2012). However, none of them are sensed tagged or aligned at word level. Chinese-English word aligned corpora are available as part of many statistical machine translation projects, but we wanted to work with a multilingual corpus, not just two languages.

Rather than translate new data, we took advantage of an existing multilingual corpus containing eight languages: English (eng), Mandarin Chinese (cmn), Japanese (jpn), Indonesian (ind), Korean, Arabic, Vietnamese and Thai (Tan and Bond, 2012). Parallel data in English, Chinese, Japanese, and Indonesian are selected for further annotation, which is composed of three genres: short stories, essays and tourism.

The Princeton Wordnet is an important resource in natural language processing, psychology, and language studies. It was developed from 1985 at Princeton University. Nouns, verbs, adjective and adverbs were grouped into synsets and linked through semantic relation (Fellbaum, 1998). We used Southeast University's Chinese Wordnet to

tag the Chinese part (SEW: Xu et al., 2008),[1] and are now in the process of switching to the Chinese Open Wordnet (COW: Wang and Bond, 2013).[2]

3 Pre-processing the Corpus

In this paper we talk only about the Chinese and English text from the short story, essay and tourism genres of the NTU-MC, although we are also cooperating with other work on tagging Indonesian and Japanese (Bond et al., 2013). The short stories are two Sherlock Holmes' Adventures (*The Adventure of the Dancing Men* and *The Adventure of the Speckled Band*), the essay is *The Cathedral and the Bazaar* (Raymond, 1999) and the tourism data is from the Singapore tourist board's web pages (Singapore Tourist Board, 2012). The corpus sizes are shown in Table 1. We show the number of sentences, words and concepts (open class words taggable with synsets).

3.1 Pre-processing with NLP Tools

For English, Freeling (Padró et al., 2010) was run with number processing, name recognition, multi-words, dates and quantities all turned off. Turning them on gave quite aggressive lemmatization: for example *a bit* in *a bit of honest money* was lemmatized to *IF_bit:1* "one bit of information". We did very minimal preprocessing: for example rewriting three hyphens `---` to mdash —. We had some problems with lemmatization of hyphenated expressions and mdashes: *white-counterpaned* which we would like to treat as two lemmas (*white* and *counterpane*) and *not—because* which should be treated as *not* and *because*. We ended up correcting many of these by hand.

For Chinese, we segmented and tagged with the Stanford NLP tools (Chang et al., 2008).[3] We did some post-processing: many punctuation marks were not recognized (such as: `[〔"()--Ⓡ{"'`), these we corrected with a script after the initial POS tagging. We also lemmatized plural-marked nouns, such as 学生+们 *xuéshēng+men* "student+s" to 学生 *xuéshēng* "student". This only occurs for 18 words.[4] The only other lemmatization we did for Chinese was for reduplicated words, where the lemma is the un-duplicated form.

Finally, we preprocess the Chinese wordnet by running it through the same segmenter, and storing the segmented forms as well.

	POS	English	Chinese
n	noun	billiard	台球 *táiqiú*
v	verb	convey	传达 *chuándá*
a	adjective	curious	奇特 *qítè*
r	adverb	finally	最后 *zuìhòu*

Table 2: Parts of Speech

3.2 Identifying Concepts

We add potential concepts as a separate layer, linked to the words (like terms in KAF: Bosma et al., 2009).

We identify concepts in two ways: words or multi-word expressions (MWEs) that are in wordnet or any single open class words not yet matched (these are tagged as **unknown**).

A word may potentially be part of multiple concepts (single and multi-word). For example *distribution* in *Gaussian or Poisson distributions* is marked as being part of ***Gaussian distribution***, ***Poisson distribution*** and ***distribution***. Concepts can be discontinuous (like *Gaussian distribution* above), we allow up to three extra words to intervene. After preliminary trials, we decided to ignore POS tags when matching words to concepts (see Section 3.4 for more discussion).

Our concepts comprise of single content words and MWE. Words fall into four major categories: n, v, a, r, following the standard wordnet structure. We show examples in Table 2.

Various heuristics were employed to make the matching flexible. For single word entries in wordnet, we match on lemmas, not using the wordnet form variants. If we can find no match for the lemma, then we try to match the surface form. All matching is done with lower-cased entries. For English, we further process entries with hyphens to produce extra forms without the hyphen: ***database*** will also match ***database*** and ***data base***.

For multiword expressions, we index them by the first token. If that matches either lemma or

[1] At the beginning of our project we tested a small sample of Chinese words by looking them up in both SEW and the Sinica Bilingual Ontological Wordnet (Huang et al., 2004) and found SEW had slightly better coverage.

[2] COW is available at `http://compling.hss.ntu.edu.sg/cow/`.

[3] We compared several free Chinese morphological analyzers and found this the most consistent.

[4] In some segmentation standards these would be two tokens, the Chinese Penn Treebank consistently treats these as one token (Xia, 2000).

Genre	English			Chinese		
	Sents	Words	Concepts	Sents	Words	Concepts
Essay	769	18,693	10,435	816	18,216	11,365
Story	1,198	22,818	11,340	1,226	23,758	12,630
Tourism	2,988	74,332	40,844	3,280	63,905	43,164

Table 1: Size of the Corpora

surface form we then continue to match the remaining tokens, allowing up to three intervening tokens. We must check both surface and lemmas to deal with cases such as ***programming language*** which is lemmatized to $program_{VV}$ $language_{NN}$. Other wordnet taggers we tested have missed many MWEs, for example, Freeling will not recognize ***look up*** in *look the word up*.

Sag et al. (2002) classified MWE into lexicalized phrases and institutionalized phrases. The former can be grouped into fixed expressions, semi-fixed expressions and syntactically-flexible expressions; the latter includes anti-collocations and collocations. All of these types are found in our corpus, as shown in Table 3.

Our matching is still imperfect. It is too loose for fixed expressions: for example, there will never be anything (except for expletives, which can also come within words) between *ad* and *hoc* (or *for* and *example*). It therefore matches many MWEs which the annotators need to discard. It is too rigid for syntactically-flexible expressions, which can have their order changed (e.g. by passivization) and thus misses some entries.

3.3 Distribution of Concepts

Table 4 shows the number of concepts in the three genres of essay, story and tourism for both Chinese and English. In each of the three subcorpora, Chinese has more concepts than English, possibly because our tagging of unknown words is less precise. We show how many are found in the wordnets (in WN: PWN for English, SEW for Chinese): the remainder are unknown open class words. The coverage is best for the stories and slightly worse for the essay (which has many technical terms, such as *developer* "one who programs computers or designs software"). It is much worse for the Singapore tourist data, which introduces many new concepts, such as *ikan bilis* "an Indonesian dish made with fried anchovies and peanuts", *mooncake* "a kind of Chinese cake eaten around the Autumn festival", *Merlion* "the statue that symbolizes Singapore" and many more. The coverage is worse for Chinese, as the wordnet is not as well developed. In addition, there are many words lexicalized in Chinese but not in English, for example, 去年 *qùnián* 'last year'. Further, there are many English foreign words in the tourism corpus, which makes the coverage even worse. Finally we show the number of concepts for which the annotators chose a single wordnet sense. Not all untagged words should be tagged however: they may be mis-identified as MWEs or open class words, named entities, mis-tokenizations or concepts not currently in wordnet.

3.4 Part of Speech Issues

For our tagging interface, we looked up wordnet using the lemma of a word. This caused problems when the word was mis-tagged giving the wrong lemma. The well-known problematic cases of present and past-participles. For example, *drawing* in *"Have you that fresh drawing?"* was tagged as VBG with lemma ***draw*** although it should have been ***drawing*** (NN). In this case, the annotators have the option of specifying the noun synset, but the first version of our tool currently did not allow them to fix the POS and lemma.[5] In general, the annotators found distinguishing between gerunds, adjective and participles hard. For example in *dancing men* (referring to pictures of little men that look as though they are dancing): should this be the noun $\textbf{\textit{dancing}}_{n:1}$ "making a series of rhythmical steps (and movements) in time to music" or the verb $\textbf{\textit{dance}}_{v:2}$ "move in a pattern; usually to musical accompaniment"? These are linked by a derivational link, so are clearly related. We decided on a general strategy and tried to make the decision process as clear as possible in the tagging guidelines, revising them with more examples. The annotators should first check if the context makes the word clearly an adjective, verb or noun, and if so pick the appropriate sense based on this. If the word is ambiguous in context, first pre-

[5]The tool now allows the annotators to change the lemma.

MWE		English	Chinese
lexicalized	fixed	point of view	不容置疑 bùróng zhìyí ‘unquestionable’
	semi-fixed	New York police bureau	乡村医生 xiāngcūn yīshēng ‘country doctor’
	syntactically-flexible	make sense	打电报 dǎ diànbào ‘send a telegram’
institutionalized	collocations	power-making	白发苍苍 bái fà cāngcāng ‘white-haired’

Table 3: Multi-word expression types

Genre	English					Chinese				
	Concepts	in WN	%	Tagged	%	Concepts	in WN	%	Tagged	%
Essay	10,435	9,588	91.9	8,607	82.5	11,365	8,620	75.8	8,773	77.2
Story	11,340	10,761	94.9	9,550	84.2	12,630	9,521	75.4	8,737	69.2
Tourism	40,844	35,979	88.1	32,990	80.8	43,164	23,699	54.9	18,663	43.2

Table 4: Distribution of Concepts and Tags

tag \ pos	n	v	a	r	x
n	12,426	970	140	129	93
v	709	7,950	14	77	19
a	1,750	2,092	1,206	836	453
r	315	390	98	4,504	191

Table 5: Confusion Matrix: POS vs Tag (Chinese)

tag \ pos	n	v	a	r	x
n	20,763	903	481	151	249
v	538	11,686	58	12	20
a	1,085	481	7,427	312	424
r	75	17	357	4,171	347

Table 6: Confusion Matrix: POS vs Tag (English)

fer an adjective if it exists, then verb, then noun. Similar guidelines were written for other confusing cases.

We show the confusion matrices of wordnet part of speech versus lemmatizer tag (simplified to the four wordnet parts of speech and other (**x**) for Chinese and English (for single word lemmas) in Tables 5 and 6 respectively. A common error was NN in English tagged as **a**: this included examples such as ***Chinese***, ***open-source*** and ***last***.

In general, the POS tagger could not be relied on. The annotators picked a different tag from the system 24.1% of the time for Chinese and 11.1% of the time for English. This shows how poorly POS taggers perform outside the domains they are trained on: real-world accuracy is between 80 and 90%.

4 Monolingual Sense Tagging

Our annotators (for both the monolingual and cross-lingual sense tagging) were undergraduate students (and recent graduates) from the linguistics and multilingual studies division at Nanyang Technological University. All were bilingual Chinese-English speakers and several had good command of Japanese. Most had experience tagging as part of the core semantics class, where a tagging exercise is used to teach about lexical semantics.

The annotators chose between existing wordnet senses based on the lemma senses or a number of meta-tags: **e, s, m, p, u**. The expectation was that after the tagging, there would be a round of wordnet extension, and then the words with new wordnet entries would be tagged once more.

Their meaning is explained below, and their distribution is given in Table 7.

- Problems in the pre-processing:

  p POS that should not be tagged (article, modal, preposition, . . .)

  e error in tokenization
  今 日*jīn rì* should be 今日*jīnrì* “today”
  three-toed should be *three - toed*

- Problems with wordnet:

  s missing sense (not in wordnet)
  I program in python “the computer language”
  COMMENT: add link to existing synset <06898352-n “programming language”

  u lemma not in wordnet but POS open class (tagged automatically)
  COMMENT: add or link to existing synset

m Multiword
(i) if the lemma is a multiword, this tag means it is not appropriate;
(ii) if the lemma is single-word, this tag means it should be part of a multiword.

The first two errors are those where the system has wrongly offered a word to be tagged, or the morphological processing has failed in some way. Because the annotators had no training in part of speech tagging, they were instructed to note the error (with a comment is possible) and these would be fixed and then re-tagged later. We have not done a full analysis, but a preliminary investigation suggests that modal auxiliaries and prepositions were the most common **p** and **e** tags. In general the annotators found it hard to distinguish between **p** and **e**: we are trying to make the guidelines clearer. The distinguishing criteria should be **e** means that the annotation should be fixed in some way, while **p** just means there is no need to annotate: the annotators had trouble making this distinction. The annotators often marked existential *there* and exclamatives (like *ah!*) as **s** "should add to wordnet", we have updated the tagging guidelines to make this clearer. Although the Penn treebank tag set does distinguish between existential and referential *there*, we check both anyway as the pos tagging is unreliable. However, to speed up tagging, because existential *there* and preposition *in* are so often **p** we pre-mark these entries as **p** before annotation. Further, although the tags do not distinguish between auxiliaries and main verbs, we found it fairly easy to identify them with simple patterns: such as, V:[***have***|***be***] V:VBG|VBN. We used these patterns to also pre-mark these entries as **p**.

Those marked with **s** and **u** are missed cases in either PWN or SEW. We can see from Table 7 that the Chinese wordnet (SEW) has many more missed senses and lemmas compared to PWN. This is one reason that we are switching to the Chinese Open Wordnet (COW) which has better accuracy and coverage (Wang and Bond, 2013).

One goal of the annotation is to improve the wordnets by adding the new words and senses, and we are working on this in parallel with the annotation. Anything tagged **s** or **u** is thus a possible new addition to wordnet. There were 1,375 such tags for English and 24,594 for Chinese. However, if we look at the distinct lemmas, then there are far fewer: 799 for English and 7,691 for Chinese. This gives us a rough estimate of how many new entries need to be created.

Tag	%	Type (Example)
p	38	Shouldn't be tagged (***ah***)
m	10	Part of known multiword (***send for***)
e	6	Wrong tokenization/lemmatization ***uptimes*** → ***uptime***
tag	14	Existing sense is ok (***idea***$_{n:1}$)
u	16	New lemma (***matter***$_{n:1}$)
s	16	New synset and lemma (***catlike***$_{n:1}$)

Table 8: Real Distribution of New Candidates

We looked at a random sample of 50 entries (tokens) marked **s** or **u** and found the situation encouraging, only 30% really required new entries. We summarize the results in Table 8, giving the correct tag, percentage, explanation and example.

As discussed above, some exclamatives, existentials and other things that should not be tagged were marked with **s**. More problematically, the annotators often marked *Watson* (Sherlock Holmes's companion) with *s*, although they had been instructed to mark proper names with **p**. Here, although technically an error, we are sympathetic: ***Sherlock Holmes*** is in wordnet, and ***John Watson*** seems prominent enough to add.

In some cases, even where they had correctly marked the multiword, they marked the single words as **s** not **m**. This is just an error. For example in (1), ***send for***$_{v:1}$ was correctly annotated, and *send* should be marked as **m** "part of multiword" rather than **s**.

(1) *They had at once sent for the doctor and for the constable.*

In some cases the lemmatizer had incorrectly lemmatized the word: *uptimes* in (2) should be lemmatized as ***uptime***, which is in wordnet "period of time when something (as a machine or factory) is functioning and available for use". This should have been tagged with **e** and the correct lemma and tag given in the comments.

(2) *[. . .] its continuous uptimes spanning months or even years.*

In a few cases (tag), we judged that an existing sense could be used. For example, in (3), the annotator wanted to tag it with ***concept***$_{n:1}$ "abstract or general idea inferred or derived from specific instances", but we judged that it was Ok as the hypernym ***idea***$_{n:1}$ "the content of cognition; the main

Genre	English					Chinese				
	p	e	s	u	m	p	e	s	u	m
Essay	552	354	258	189	418	202	40	178	1,846	167
Story	825	186	185	12	495	459	300	1,263	1,041	524
Tourism	1,630	954	286	445	2,278	937	431	2,769	17,497	494

Table 7: Distribution of Meta-Tags

thing you are thinking about" which has as its example: *it was not a good idea*. In some cases, we thought that the definition should be made clearer (often less dogmatic) in order to make the scope of the sense wider. For example in (4), wordnet has ***backer***$_{n:1}$ "invests in a theatrical production", as a hyponym of ***patron***$_{n:1}$. We feel this could be expanded to "a person who invests in something, such as a theatrical production", avoiding the construction of a new sense.

(3) *Though fetchpop had some good original ideas in it (such as its background-daemon mode)*

(4) *[. . .] the open-source idea has scored successes and found backers elsewhere.*

Finally, there were some genuinely new senses. *The Cathedral and the Bazaar* made many references to *developers* and *co-developers*. ***developer*** is almost certainly derived from ***develop***$_{v:1}$ "make something new, such as a product or a mental or artistic creation" and ***co-developer*** from there. Some were rare uses of existing words as in (5), where *matter* meaning ***measure***$_{n:1}$ "how much there is or how many there are of something that you can quantify" is an established if old-fashioned use, some were common extensions of existing entries, as in (6), where *toolkit* refers to the skills a person possesses rather than the physical ***tool kit***$_{n:1}$ "a set of carpenter's tools", and should be a synonym for ***bag of tricks***$_{n:1}$ "supply of ways of accomplishing something".

(5) *[. . .] my people have been at Riding Thorpe for a matter of five centuries [. . .]*

(6) *[. . .] it increases the probability that someone's toolkit will be matched to the problem [. . .]*

5 Cross-lingual Annotation

For the second round of annotation, instead of going over the monolingual texts again, we decided to look at the sense annotation in the translation context.

For each sentence, we automatically linked words with either: the same concept (=); if still unlinked then a matching hypernym or hyponym ($<$, $>$); if still unlinked then the same lemma (this was useful even between English and Chinese as technical terms (such as ***Linux***) were often left in the Latin alphabet). We did not use word-to-word tags in the tagging because (i) they were unavailable and (ii) we already had the monolingual tags on each side, so we did not need to project the tags. In future work, we would like to investigate the use of word-links (following the lead of Bentivogli and Pianta (2005)).

The annotators then went through sentence-pair by sentence-pair and (i) checked existing links then (ii) tried to link unlinked concepts. They categorized links into the six types shown in Table 9. The annotators were instructed not to overthink the decision as to link-type: we can recalculate the distinctions using the wordnet structure.

This annotation has only been completed for the Essay and Story genres, we show the numbers of links, and the total number of taggable concepts, in Table 10. The proportion of things linked is very low: 61% for the stories and 39% for the essay. We have automatically calculated the types of links: if the tag is exactly the same, then =; hyponyms and hypernyms are shown with $<$ and $>$; derivationally related forms and pertainyms found in wordnet with **d**; other linked tags with different parts-of-speech with **D**; holonyms with **m**; meronyms with **M**; antonyms found in wordnet with **!**; those the annotator marked as antonyms but we could not find in wordnet with **#** and everything else with $\sim$. The large number of part-of-speech mismatches suggests that we still do not have all the cross part-of-speech links in wordnet that we should.

An example of why things remain unlinked is shown in (7): concepts are marked with subscripts, linked concepts have the same subscript and are upper case. ***way*** and ***question*** can be linked, but ***put*** and ***answer*** can not, even though the transla-

Symbol	Explanation	English	Chinese
=	same synset	about	大约 *dàyuē* “about”
<	hyponym	armchair	椅子 *yǐzi* “chair”
>	hypernym	body	遗体 *yítǐ* “remains”
∼	lexically linked	absorb	全神贯注 *quánshénguànzhù* “with breathless interest”
≈	pragmatically linked	absurdly	太 *tài* “excessively”
!	antonym	easy	难 *nán* “difficult”
#	weak antonym	miss	打中 *dǎ zhòng* “hit”

Table 9: Link Types with Examples

Link	Story		Essay	
	#	%	#	%
=	2,642	41.7	2,155	48.9
<	107	1.7	31	0.7
>	205	3.2	123	2.8
∼	2184	34.5	1464	33.2
d	166	2.6	72	1.6
D	1,149	18.1	624	14.2
m	16	0.3	1	0.0
M	15	0.2	5	0.1
!	2	0.0	0	0.0
#	23	0.4	7	0.2
Total	6,336	100.0	4,407	100.0
Concepts	10,435		11,340	

Table 10: Number of Links

tion clearly has the same meaning. In general *The Cathedral and the Bazaar* had more complicated prose than the stories, and the translations were less well aligned. Arguably, ***put*** could be linked to wèn with ≈ but the annotator did not do so.

(7) Put$_a$ that way,$_B$ the question$_c$ answers$_D$ itself.

这样$_B$ 一 问$_e$， 答案$_D$ 自明$_f$。
zhèyàng yī wèn, dá'àn zìmíng.
like this one ask, answer self-evident

“Asking like this, the answer is self-evident.”

Often, there were differences in lexicalization that made the question of what to link difficult. For example in (8), 前额 *qián'é* “forehead” is lexicalized, and it matches to a unit that is not in PWN, “the front of ones brain”. This is almost certainly not lexicalized in English. So we end up linking *qián'é* to ***brain*** with ≈ and then ***front*** has nothing to link to. We need to be able to link words to phrases without necessarily adding the phrases to the wordnets.

(8) The bullet had passed through the front of her brain.

子弹 是 从 她的 前额 打 进去 的。
Zǐdàn shì cóng tā de qián'é dǎ jìnqù de.
bullet is from her forehead shoot enter
“The bullet was shot in from her forehead”

6 Discussion and Further Work

We have been gradually improving the tagging guidelines as we continue with the annotation, and will make these available online along with the corpus.[6] In particular, we are adding more examples for each case. We benefited from the cheat sheet and guidelines produced for the Gloss Corpus (Fellbaum pc.) and hope our guidelines can help other people. With this in mind, we are trying to keep separate, as far as possible, tool-specific procedures and general tagging guidelines.

Many of the unknown words, especially for our first attempt, were in fact words that are in wordnet with minor typographical variations: for example *tool kit* in wordnet as ***toolkit***.[7] We have added various heuristics to improve the look up within wordnet. We also started to work on improving the tokenization, but decided this was too large a task. Instead, we are looking at exploiting a more semantically aware tokenizer (Dridan and Oepen, 2012). Similarly for Chinese, we are comparing a wider variety of tokenizers. One reviewer suggested that there are more accurate proprietary pos taggers and segmenters available for Chinese. Unfortunately, the fact that they are not freely available means that we cannot test them to see if they are better. Our experience with English,

[6] The corpus and guidelines are available at `http://compling.hss.ntu.edu.sg/ntumc/`.

[7] Although not with the desired sense.

where we have more experience with state-of-the-art systems is that (i) they do not do well with out-of-domain data (a well-known failing) and (ii) they often do not mark distinctions important for the sense tagging (for example, the difference between main and auxiliary verbs). We therefore prefer to work with open-source systems that we can evaluate and potentially improve.

In this paper, we mainly discuss a breadth first approach, where we are trying to increase the coverage uniformly to cover all words. We do not report on the inter-annotator agreement, as the first rounds of tagging (which we report on here) are not the final annotation: all tags are checked once more as we do the cross-lingual annotation, and it is too expensive to do this multiple times.

We are also using the corpora as a test-bed to look at individual phenomena of interest in detail, including the use of Chinese traditional idiomatic expressions (成语 *chéngyǔ*), English possessive idioms (*X looses X's head*) and the differences in pronoun distribution across languages.

7 Conclusions

This paper presents preliminary results from an ongoing project to construct large-scale sense-tagged parallel corpora. The annotation scheme is divided into two phrases: monolingual sense annotation and multilingual concept alignment. We look at some of the issues raised for Chinese and English annotation of text in three genres. All annotated corpora will be made freely available, in addition, the changes made to the wordnets will be released through the individual wordnet projects.

Acknowledgments

This research was supported in part by the MOE Tier 1 grant *Shifted in Translation — An Empirical Study of Meaning Change across Languages*. We would like to thank our annotators: En Jia Chee, Eshley Gao, Jeanette Tan, Hui Ting, Wanxuan Wang, and Hazel Wen. Finally, would like to thank Huizhen Wang for her many helpful discussions.

References

Luisa Bentivogli and Emanuele Pianta. 2005. Exploiting parallel texts in the creation of multilingual semantically annotated resources: the multisemcor corpus. *Natural Language Engineering*, 11(3):247–261.

Francis Bond, Timothy Baldwin, Richard Fothergill, and Kiyotaka Uchimoto. 2012. Japanese SemCor: A sense-tagged corpus of Japanese. In *Proceedings of the 6th Global WordNet Conference (GWC 2012)*, pages 56–63. Matsue.

Francis Bond and Kyonghee Paik. 2012. A survey of wordnets and their licenses. In *Proceedings of the 6th Global WordNet Conference (GWC 2012)*. Matsue. 64–71.

Francis Bond, Shan Wang, Eshley Huini Gao, Hazel Shuwen Mok, and Jeanette Yiwen Tan. 2013. Developing parallel sense-tagged corpora with wordnets. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse (LAW 2013)*, pages 149–158. Sofia. URL `http://www.aclweb.org/anthology/W13-2319`.

Wauter Bosma, Piek Vossen, Aitor Soroa, German Rigau, Maurizio Tesconi, Andrea Marchetti, Monica Monachini, and Carlo Aliprandi. 2009. KAF: a generic semantic annotation format. In *Proceedings of the 5th International Conference on Generative Approaches to the Lexicon (GL 2009)*. Pisa.

Baobao Chang. 2004. Chinese-English parallel corpus construction and its application. In *Proceedings of The 18th Pacific Asia Conference on Language, Information and Computation*, pages 283–290. Waseda University, Tokyo.

Baobao Chang, Weidong Zhan, and huarui Zhang. 2003. The construction and management of a bilingual corpus used for Chinese-English machine translation (面向汉英机器翻译的双语语料库的建设及其管理). *Terminology Standardization & Information Technology (*术语标准化与信息技术*)*, (1):28–31.

Pi-Chuan Chang, Michel Galley, and Christopher D. Manning. 2008. Optimizing Chinese word segmentation for machine translation performance. In *ACL Third Workshop on Statistical Machine Translation*.

Yidong Chen, Xiaodong Shi, and Changle Zhou. 2006. Research on filtering parallel corpus: A ranking model (平行语料库处理初探:一种排序模型). *Journal Of Chinese Information Processing*, 20(z1).

Yidong Chen, Xiaodong Shi, Changle Zhou, and Qing-Yang Hong. 2005. A model for ranking sentence pairs in parallel corpora. In *Machine Learning and Cybernetics, 2005. Proceedings of 2005 International Conference on*, volume 6, pages 3820–3823. IEEE.

Rebecca Dridan and Stephan Oepen. 2012. Tokenization: Returning to a long solved problem: A survey, contrastive experiment, recommendations, and toolkit. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 378–382. Association for Computational Linguistics, Jeju Island, Korea. URL `http://www.aclweb.org/anthology/P12-2074`.

Christine Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Chu-Ren Huang, RU-Yng Chang, and Shiang-Bin Lee. 2004. Sinica BOW: integrating bilingual WordNet and SUMO ontology. In *Proceedings of the Fourth International Language Resources and Evaluation (LREC 2004)*, pages 1553–1556.

Shari Landes, Claudia Leacock, and Christiane Fellbaum. 1998. Building semantic concordances. In Fellbaum (1998), chapter 8, pages 199–216.

Monica Lupu, Diana Trandabat, and Maria Husarciuc. 2005. A Romanian semcor aligned to the English and Italian multisemcor. In *Proceedings 1st ROMANCE FrameNet Workshop at EUROLAN 2005 Summer School*, pages 20–27. EUROLAN, Cluj-Napoca, Romania.

Lluís Padró, Miquel Collado, Samuel Reese, Marina Lloberes, and Irene Castellón. 2010. Freeling 2.1: Five years of open-source language processing tools. In *Proceedings of 7th Language Resources and Evaluation Conference (LREC 2010)*. La Valletta. URL `http://nlp.lsi.upc.edu/freeling`.

Tommaso Petrolito and Francis Bond. 2014. A survey of wordnet annotated corpora. In *Proceedings of the 7th Global WordNet Conference (GWC 2014)*. Tartu. (this volume).

Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. 2002. Multiwordnet: Developing an aligned multilingual database. In *In Proceedings of the First International Conference on Global WordNet*, pages 293–302. Mysore, India.

Eric S. Raymond. 1999. *The Cathedral & the Bazaar*. O'Reilly.

Ivan Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In Alexander Gelbuk, editor, *Computational Linguistics and Intelligent Text Processing: Third International Conference: CICLing-2002*, pages 1–15. Springer-Verlag, Hiedelberg/Berlin.

Singapore Tourist Board. 2012. Your Singapore. Online: `http://www.yoursingapore.com`. [Accessed 2012].

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

Lucy Vanderwende and Aral Menezes. 2005. The empirical revolution in natural language processing. In *4th International Conference on Natural Language Processing (ICON 2005)*, pages 7–8. (Invited Talk).

Kefei Wang. 2012. On the design and construction of the super-large-scale China English-Chinese parallel corpus (CECPC) (中国英汉平行语料库的设计与研制). *Foreign Languages in China (*中国外语*)*, pages 23–27.

Shan Wang and Francis Bond. 2013. Building a Chinese wordnet: Starting from core synsets. In *Proceedings of the 11th Workshop on Asian Language Resources*. Nagoya.

Fei Xia. 2000. The segmentation guidelines for the Penn Chinese Treebank (3.0). Technical Report IRCS-00-06, University of Pennsylvania Institute for Research in Cognitive Science.

Renjie Xu, Zhiqiang Gao, Yuzhong Qu, and Zhisheng Huang. 2008. An integrated approach for automatic construction of bilingual Chinese-English WordNet. In *3rd Asian Semantic Web Conference (ASWC 2008)*, pages 302–341.