Developing Parallel Sense-tagged Corpora with Wordnets

Francis Bond, Shan Wang,
Eshley Huini Gao, Hazel Shuwen Mok, Jeanette Yiwen Tan
Linguistics and Multilingual Studies, Nanyang Technological University
bond@ieee.org

Abstract

Semantically annotated corpora play an important role in natural language processing. This paper presents the results of a pilot study on building a sense-tagged parallel corpus, part of ongoing construction of aligned corpora for four languages (English, Chinese, Japanese, and Indonesian) in four domains (story, essay, news, and tourism) from the NTU-Multilingual Corpus. Each subcorpus is first sense-tagged using a wordnet and then these synsets are linked. Upon the completion of this project, all annotated corpora will be made freely available. The multilingual corpora are designed to not only provide data for NLP tasks like machine translation, but also to contribute to the study of translation shift and bilingual lexicography as well as the improvement of monolingual wordnets.

1 Introduction

Large scale annotated corpora play an essential role in natural language processing (NLP). Over the years with the efforts of the community part-of-speech tagged corpora have achieved high quality and are widely available. In comparison, due to the complexity of semantic annotation, sense tagged parallel corpora develop slowly. However, the growing demands in more complicated NLP applications such as information retrieval, machine translation, and text summarization suggest that such corpora are in great need. This trend is reflected in the construction of two types of corpora: (i) parallel corpora: FuSe (Cyrus, 2006), SMULTRON (Volk et al., 2010), CroCo (Čulo et al., 2008), German-English parallel corpus (Padó and Erk, 2010), Europarl corpus (Koehn, 2005), and OPUS (Nygaard and Tiedemann, 2003; Tiedemann and Nygaard, 2004; Tiedemann, 2009, 2012) and (ii) sense-tagged monolingual corpora: English corpora such as Semcor (Landes et al., 1998); Chinese corpora, such as the crime domain of Sinica Corpus 3.0 (Wee and Mun, 1999), 1 million word corpus of People's Daily (Li et al., 2003), three months' China Daily (Wu et al., 2006); Japanese corpora, such as Hinoki Corpus (Bond et al., 2008) and Japanese SemCor (Bond et al., 2012) and Dutch Corpora such as the Groningen Meaning Bank (Basile et al., 2012). Nevertheless, almost no parallel corpora are sense-tagged. With the exception of corpora based on translations of SemCor (Bentivogli et al., 2004; Bond et al., 2012) sense-tagged corpora are almost always monolingual.

This paper describes ongoing work on the construction of a sense-tagged parallel corpus. It comprises four languages (English, Chinese, Japanese, and Indonesian) in four domains (story, essay, news, and tourism), taking texts from the NTU-Multilingual Corpus (Tan and Bond, 2012). For these subcorpora we first sense tag each text monolingually and then link the concepts across the languages. The links themselves are typed and tell us something of the nature of the translation. The annotators are primarily multilingual students from the division of linguistics and multilingual studies (NTU) with extensive training. In this paper we introduce the planned corpus annotation and report on the results of a completed pilot: annotation and linking of one short story: *The Adventure of the Dancing Men* in Chinese, English and Japanese. All concepts that could be were aligned and their alignments annotated.

The paper is structured as follows. Section 2 reviews existing parallel corpora and sense tagged corpora that have been built. Section 3 introduces the resources that we use in our annotation project. The annotation scheme for the multilingual corpora is laid out in Section 4. In Section 5 we report

Proceedings of the 7th Linguistic Annotation Workshop & Interoperability with Discourse, pages 149–158,
Sofia, Bulgaria, August 8-9, 2013. ©2013 Association for Computational Linguistics

in detail the results of our pilot study. Section 6 presents our discussion and future work.

2 Related Work

In recent years, with the maturity of part-of-speech (POS) tagging, more attention has been paid to the practice of getting parallel corpora and sense-tagged corpora to promote NLP.

2.1 Parallel Corpora

Several research projects have reported annotated parallel corpora. Among the first major efforts in this direction is FuSe (Cyrus, 2006), an English-German parallel corpus extracted from the EUROPARL corpus (Koehn, 2005). Parallel sentences were first annotated mono-lingually with POS tags and lemmas; related predicates (e.g. a verb and its nominalization are then linked). SMULTRON (Volk et al., 2010) is a parallel treebank of 2,500 sentences from different genres: a novel, economy texts from several sources, a user manual and mountaineering reports. Most of the corpus is German-English-Swedish parallel text, with additional texts in French and Spanish. CroCo (Čulo et al., 2008) is a German-English parallel and comparable corpus of a dozen texts from eight genres, totaling approximately 1,000,000 words. Each sentence is annotated with phrase structures and grammatical functions, and words, chunks and phrases are aligned across parallel sentences. This resource is limited to two languages, English and German, and is not systematically linked to any semantic resource. Padó and Erk (2010) have conducted a study of translation shifts on a German-English parallel corpus of 1,000 sentences from EUROPARL annotated with semantic frames from FrameNet and word alignments. Their aim was to measure the feasibility of frame annotation projection across languages.

The above corpora have been used for studying translation shift. Plain text parallel corpora are also widely used in NLP. The Europarl corpus collected the parallel text in 11 official languages of the European Union (i.e. Danish, German, Greek, English, Spanish, Finnish, French, Italian, Dutch, Portuguese, and Swedish) from proceedings of the European Parliament. Each language is composed of about 30 million words (Koehn, 2005). Newer versions have even more languages. OPUS v0.1 contains the documentation of the office package OpenOffice with a collection of 2,014 files in English and five translated texts, namely, French, Spanish, Swedish, German and Japanese. This corpus consists of 2.6 million words (Nygaard and Tiedemann, 2003; Tiedemann and Nygaard, 2004; Tiedemann, 2012). However, when we examined the Japanese text, we found the translations are often from different versions of the software and not synchronized very well.

2.2 Sense Tagged Corpora

Surprisingly few languages have sense tagged corpora. In English, Semcor was built by annotating texts from the Brown Corpus using the sense inventory of WordNet 1.6 (Fellbaum, 1998) and has been mapped to subsequent WordNet versions (Landes et al., 1998). The Defense Science Organization (DSO) corpus annotated the 191 most frequent and ambiguous nouns and verbs from the combined Brown Corpus and Wall Street Journal Corpus using WordNet 1.5. The 191 words comprise of 70 verbs with an average sense number of 12 and 121 nouns with an average sense number of 7.8. The verbs and nouns respectively account for approximately 20% of all verbs and nouns in any unrestricted English text (Ng and Lee, 1996). The WordNet Gloss Disambiguation Project uses Princeton WordNet 3.0 (PWN) to disambiguate its own definitions and examples.[1]

In Chinese, Wee and Mun (1999) reported the annotation of a subset of Sinica Corpus 3.0 using HowNet. The texts are news covering the crime domain with 30,000 words. Li et al. (2003) annotated the semantic knowledge of a 1 million word corpus from *People's Daily* with dependency grammar. The corpus include domains such as politics, economy, science, and sports. (Wu et al., 2006) described the sense tagged corpus of Peking University. They annotated three months of the People's Daily using the Semantic Knowledge-base of Contemporary Chinese (SKCC)[2]. SKCC describes the features of a word through attribute-value pairs, which incorporates distributional information.

In Japanese, the Hinoki Corpus annotated 9,835 headwords with multiple senses in Lexeed: a Japanese semantic lexicon (Kasahara et al., 2004) To measure the conincidence of tags and difficulty degree in identifying senses, each word was annotated by 5 annotators (Bond et al., 2006).

[1] http://wordnet.princeton.edu/glosstag.shtml

[2] http://ccl.pku.edu.cn/ccl_sem_dict/

We only know of two multi-lingual sense-tagged corpora. One is MultiSemCor, which is an English/Italian parallel corpus created based on SemCor (Landes et al., 1998). MultiSemCor is made of 116 English texts taken from SemCor with their corresponding 116 Italian translations. There are 258,499 English tokens and 267,607 Italian tokens. The texts are all aligned at the word level and content words are annotated with POS, lemma, and word senses. It has 119,802 English words semantically annotated from SemCor and 92,820 Italian words are annotated with senses automatically transferred from English (Bentivogli et al., 2004). Japanese SemCor is another translation of the English SemCor, whose senses are projected across from English. It takes the same texts in MultiSemCor and translates them into Japanese. 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). The low annotation rate compared to MultiSemCor reflects both a lack of coverage in the Japanese wordnet and the greater typological difference.

Though many efforts have been devoted to the construction of sense tagged corpora, the majority of the existing corpora are monolingual, relatively small in scale and not all freely available. To the best of our knowledge, no large scale sense-tagged parallel corpus for Asian languages exists. Our project will fill this gap.

3 Resources

This section introduces the wordnets and corpora we are using for the annotation task.

3.1 Wordnets

Princeton WordNet (PWN) is an English lexical database created at the Cognitive Science Laboratory of Princeton University. It 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. (Miller, 1998; Fellbaum, 1998). The version we use in this study is 3.0.

A number of wordnets in various languages have been built based on and linked to PWN. The Open Multilingual Wordnet (OMW) project[3] currently provides 22 wordnets (Bond and Paik, 2012; Bond and Foster, 2013). The Japanese and Indonesian wordnets in our project are from OMW provided by the creators (Isahara et al., 2008, Nurril Hirfana et al., 2011).

The Chinese wordnet we use is a heavily revised version of the one developed by Southeast University (Xu et al., 2008). This was automatically constructed from bilingual resources with minimal hand-checking. It has limited coverage and is somewhat noisy, we have been revising it and use this revised version for our annotation.

3.2 Multilingual Corpus

The NTU-multilingual corpus (NTU-MC) is compiled at Nanyang Technological University. It contains eight languages: English (eng), Mandarin Chinese (cmn), Japanese (jpn), Indonesian (ind), Korean, Arabic, Vietnamese and Thai (Tan and Bond, 2012). We selected parallel data for English, Chinese, Japanese, and Indonesian from NTU-MC to annotate. The data are from four genres, namely, short story (two Sherlock Holmes' Adventures), essay (Raymond, 1999), news (Kurohashi and Nagao, 2003) and tourism (Singapore Tourist Board, 2012). The corpus sizes are shown in Table 1. We show the number of words and concepts (open class words tagged with synsets) only for English, the other languages are comparable in size.

4 Annotation Scheme for Multilingual Corpora

The annotation task is divided into two phases: monolingual sense annotation and multilingual concept alignment.

4.1 Monolingual Sense Annotation

First, the Chinese, Japanese and Indonesian corpora were automatically tokenized and tagged with parts-of-speech. Secondly, concepts were tagged with candidate synsets, with multiword expressions allowing a skip of up to 3 words. Any match with a wordnet entry was considered a potential concept.

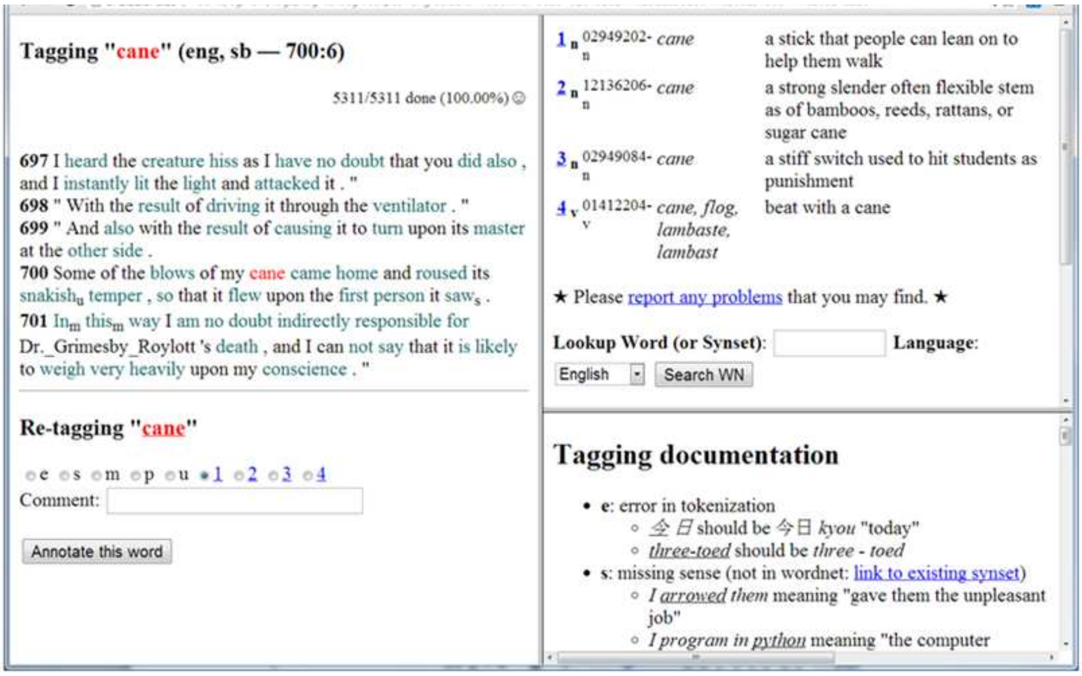
These were then shown to annotators to either select the appropriate synset, or point out a problem. The interface for doing sense annotation is shown in Figure 1.

In Figure 1, the concepts to be annotated are shown as red and underlined. When clicking on

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

Genre	Text	Sentences				Words	Concepts
		Eng	Cmn	Jpn	Ind	Eng	Eng
Story	The Adventure of the Dancing Men	599	606	698	–	11,200	5,300
	The Adventure of the Speckled Band	599	612	702	–	10,600	4,700
Essay	The Cathedral and the Bazaar	769	750	773	–	18,700	8,800
News	Mainichi News	2,138	2,138	2,138	–	55,000	23,200
Tourism	Your Singapore (web site)	2,988	2,332	2,723	2,197	74,300	32,600

Table 1: Multilingual corpus size



Tagging "cane" (eng, sb — 700:6)

5311/5311 done (100.00%) ☺

697 I heard the creature hiss as I have no doubt that you did also , and I instantly lit the light and attacked it . "
698 " With the result of driving it through the ventilator . "
699 " And also with the result of causing it to turn upon its master at the other side .
700 Some of the blows of my cane came home and roused its snakish$_u$ temper , so that it flew upon the first person it saw$_s$.
701 In$_m$ this$_m$ way I am no doubt indirectly responsible for Dr._Grimesby_Roylott 's death , and I can not say that it is likely to weigh very heavily upon my conscience . "

Re-tagging "cane"

○ e ○ s ○ m ○ p ○ u ● 1 ○ 2 ○ 3 ○ 4
Comment:

1 n	02949202-n	*cane*	a stick that people can lean on to help them walk
2 n	12136206-n	*cane*	a strong slender often flexible stem as of bamboos, reeds, rattans, or sugar cane
3 n	02949084-n	*cane*	a stiff switch used to hit students as punishment
4 v	01412204-v	*cane, flog, lambaste, lambast*	beat with a cane

★ Please report any problems that you may find. ★

Lookup Word (or Synset): Language: English

Tagging documentation

- e: error in tokenization
  - 今 日 should be 今日 *kyou* "today"
  - *three-toed* should be *three - toed*
- s: missing sense (not in wordnet: link to existing synset)
  - *I arrowed them* meaning "gave them the unpleasant job"
  - *I program in python* meaning "the computer

Figure 1: Tagging the sense of *cane*.

a concept, its WordNet senses appear to the right of a screen. The annotator chooses between these senses or a number of meta-tags: **e, s, m, p, u**. Their meaning is explained below.

e error in tokenization
今 日 should be 今日
three-toed should be *three - toed*

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

m bad 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

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

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

Missing senses in the wordnets were a major issue when tagging, especially for Chinese and Japanese. We allowed the annotators to add candidate new senses in the comments; but these were not made immediately available in the tagging interface. As almost a third of the senses were missing in Chinese and Japanese, this slowed the annotators down considerably.

Our guidelines for adding new concepts or linking words to existing cover four cases:

= When a word is a synonym of an existing word, add =synset to the comment: e.g. for laidback, it is a synonym of `02408011-a` "laid-back, mellow", so we add `=02408011-a` to the comment for laidback.

< When a word is a hyponym/instance of

an existing word, mark it with <synset: For example, *python* is a hyponym of `06898352-n` ***programming language***, so we add `<06898352-n` to *python*

! Mark antonyms with !synset.

∼ If you cannot come up with a more specific relationship, just say the word is related in some way to an existing synset with ∼synset; and add more detail in the comment.

Finally, we have added more options for the annotators: **prn** (pronouns) and seven kinds of named entities: **org** (organization); **loc** (location); **per** (person); **dat** (date/time); **num** (number); **oth** (other) and the super type **nam** (name). These basically follow Landes et al. (1998, p207), with the addition of number, date/time and name. Name is used when automatically tagging, it should be specialized later, but is useful to have when aligning. Pronouns include both personal and indefinite-pronouns. Pronouns are not linked to their monolingual antecedents, just made available for cross-lingual linking.

4.2 Multilingual Concept Alignment

We looked at bitexts: the translated text and its source (in this case English). Sentences were already aligned as part of the NTU-Multilingual Corpus. The initial alignment was done automatically: concepts that are tagged with the same synset or related synsets (one level of hyponymy) are directly linked. Then the sentence pairs are presented to the annotator, using the interface shown in Figure 2.

In the alignment interface, when you hover over a concept, its definition from PWN is shown in a pop-up window at the top. Clicking concepts in one language and then the other produces a candidate alignment: the annotator then choses the kind of alignment. After concepts are aligned they are shown in the same color. Both *bell* and 门铃 *ménlíng* "door bell" have the same synset, so they are linked with =. Similarly, *Watson* and 华生 *Huáshēng* "Watson" refer to the same person, so they are also connected with =. However, *ring* in the English sentence is a noun while the corresponding Chinese word 响 *xiǎng* "ring" is a verb; so they are linked with the weaker type ∼.

We found three issues came up a lot during the annotation: (i) Monolingual tag errors; (ii) multiword expression not tagged; (iii) Pronouns not tagged.

(i) In some cases, the monolingual tag was not the best choice. Looking at the tagging in both languages often made it easier to choose between similar monolingual tags, and the annotators found themselves wanting to retag a number of entries.

(ii) It was especially common for it to become clear that things should have been tagged as multiword expressions. Consider *kuchi-wo hiraku* "speak" in (1).

(1) Said he suddenly

a.

ホームズ	が	突然	口	を	開く
ho-muzu	ga	totsuzen	kuchi	wo	hiraku
Holmes	NOM	suddenly	mouth	ACC	open

"Holmes opens his mouth suddenly"

This was originally tagged as "open mouth" but in fact it is a multiword expression with the meaning "say", and is parallel in meaning to the original English text. As this concept is lexicalized, the annotator grouped the words together and tagged the new concept to the synset `00941990-v` "express in speech". The concepts were then linked together with ˜. It is hard for the monolingual annotator to consistently notice such multiword expressions: however, the translation makes them more salient.

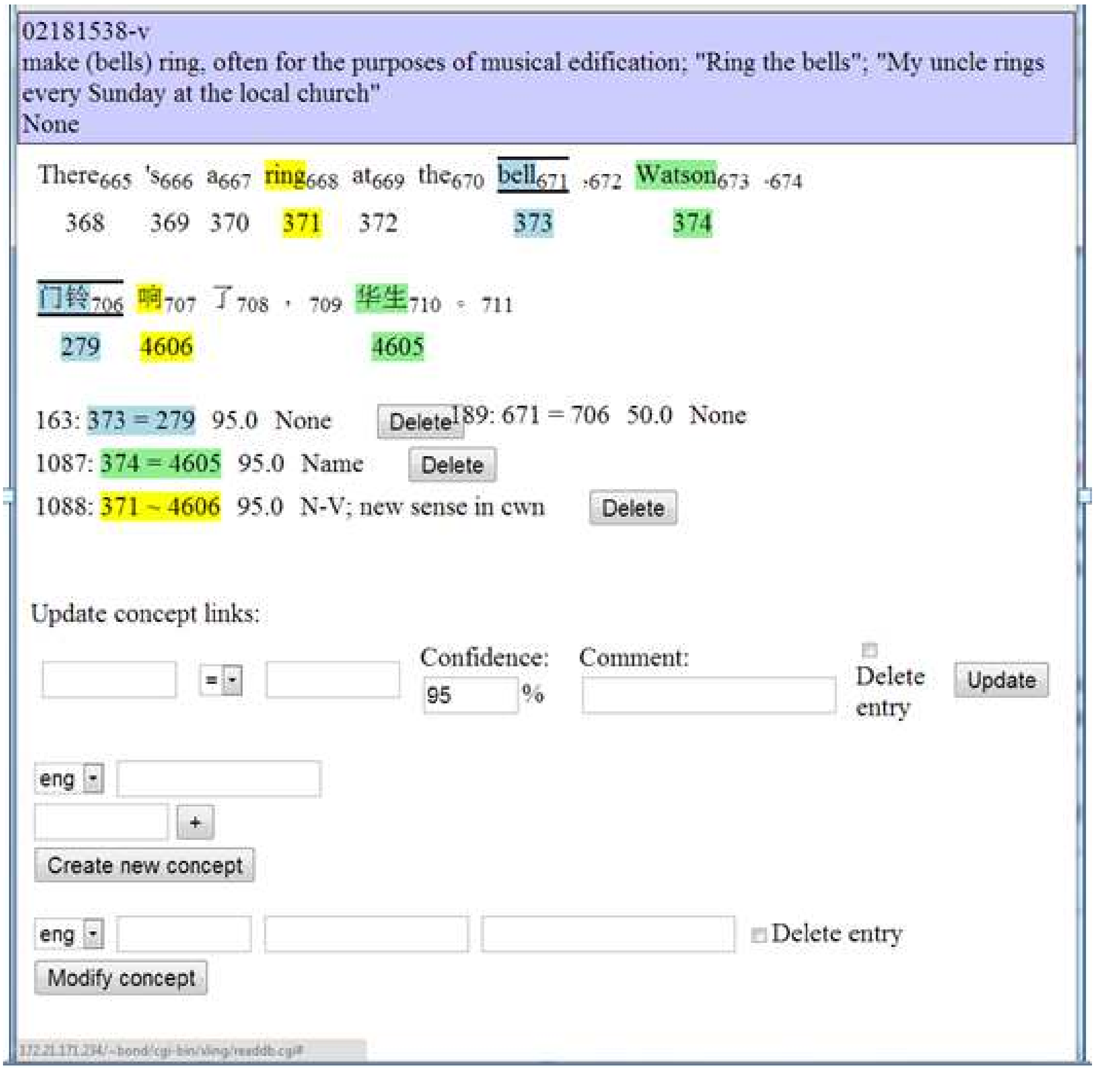
(iii) It was often the case that an open class word in one language would link to a closed class word in the other, especially to a pronoun. We see this in (1) where *he* in English links to *ho-muzu* "Holmes" in Japanese. In order to capture these correspondences, we allowed the annotator to also tag named entities, pronouns and interrogatives. From now on we will tag these as part of the initial monolingual alignment.

We tagged the links between concepts with the types shown in Table 2.

5 Pilot Study Results

A pilot study was conducted using the first story text: *The Adventure of the Dancing Men*, a Sherlock Holmes short story (Conan Doyle, 1905). The Japanese version was translated by Otokichi Mikami and Yu Okubu;[4] we got the translated version of Chinese from a website which later disappeared. Using English text as the source language, the Japanese and Chinese texts were aligned and

[4] `http://www.aozora.gr.jp/cards/000009/card50713.html`



02181538-v
make (bells) ring, often for the purposes of musical edification; "Ring the bells"; "My uncle rings every Sunday at the local church"
None

There$_{665}$'s$_{666}$ a$_{667}$ ring$_{668}$ at$_{669}$ the$_{670}$ bell$_{671}$.$_{672}$ Watson$_{673}$.$_{674}$
368 369 370 371 372 373 374

门铃$_{706}$ 响$_{707}$ 了$_{708}$ ，$_{709}$ 华生$_{710}$ 。$_{711}$
279 4606 4605

163: 373 = 279 95.0 None
189: 671 = 706 50.0 None
1087: 374 = 4605 95.0 Name
1088: 371 ~ 4606 95.0 N-V; new sense in cwn

Update concept links:

Confidence: 95 %

Figure 2: Interface for aligning concepts.

manually sense-tagged with reference to their respective wordnets. The number of words and concepts for each language is shown in Table 3.

	English	Chinese	Japanese
Sentences	599	680	698
Words	11,198	11,325	13,483
Concepts	5,267	4,558	4,561

Excluding candidate concepts rejected by the annotators.

Table 3: Concepts in Dancing Men

The relationships between words were tagged using the symbols in in Table 2. The difficult cases are similar relation and translation equivalent relation. Due to translation styles and language divergence, some concepts with related meaning cannot be directly linked. We give examples in (2) through (4).

(2) “How on earth do you know that?” I asked.

a. 「いったい 、 どうして その こと=を
「ittai 、 doushite sono koto=wo
“ on+earth , why that thing=ACC
？」 と 私=は 聞き=返す
？」 to watashi=wa kiki=kaesu
? ” QUOT me=TOP ask=return

“Why on earth do you know that thing?” I ask in return.

In (2), compared to *ask* in English, the Japanese *kikikaesu* has the additional meaning of “in return”: it is a hyponym. We marked their relation as ∼ (similar in meaning).

We introduced a new class ≈ to indicate combinations of words or phrases that are translation equivalents of the original source but are not lexicalized enough to be linked in the wordnet. One example is shown in (3).

(3) be content with my word

	Type	Example
=	same concept	*say* ↔言う *iu* “say”
⊃	hypernym	*wash* ↔洗い落とす *araiotosu* “wash out”
$\supset^2$	2nd level	*dog* ↔ 動物 *doubutsu* “animal”
⊂	hyponym	*sunlight* ↔光 *hikari* “light”
$\subset^n$	nth level	
∼	similar	*notebook* ↔ メモ帳 *memochou* “notepad”
		dull$_a$ ↔ くすむ *kusumu* “darken”
≈	equivalent	*be content with my word* ↔
		わたくし の 言葉 を 信じ-て “believe in my words”
!	antonym	*hot* ↔寒く=ない *samu=ku nai* “not cold”
#	weak ant.	*not propose to invest* ↔
		思いとどまる *omoi=todomaru* “hold back”

Table 2: Translation Equivalence Types

a. わたくし=の 言葉=を 信じ=て
watakushi=no kotoba=wo shinji=te
me=of word=ACC believe=ing
“believe in my words”

In this case *shinjite* “believe” is being used to convey the same pragmatic meaning as *content with* but they are not close enough in meaning that we want to link them in the lexicon.

(4) shows some further issues in non-direct translation.

(4) I am sure that I shall say$_h$ no$_i$thing$_j$ of the kind$_k$.

a. いやいや 、そんな こと は
iyaiya , sonna$_k$ koto$_j$ wa
by+no+means , that+kind$_k$+of thing$_j$ TOP
言わ-ん よ
iwa$_h$-n$_i$ yo
say$_h$-NEG$_i$ yo
“no no, I will not say that kind of thing”

Say$_h$ no$_i$thing$_j$ of the kind$_k$ becomes roughly “not$_i$ say$_h$ that kind$_k$ of thing$_j$”. All the elements are there, but they are combined in quite a different structure and some semantic decomposition would be needed to link them. Chinese and Japanese do not use negation inside the NP, so this kind of difference is common. Tagging was made more complicated by the fact that determiners are not part of wordnet, so it is not clear which parts of the expression should be tagged.

Though there are many difficult cases, the most common case was for two concepts to share the same synset and be directly connected. For example, *notebook* is tagged with the synset `06415419-n`, defined as “a book with blank pages for recording notes or memoranda”. In the Japanese version, this concept is translated into 備忘録 *bibouroku* “notebook”, with exactly the same synset (`06415419-n`). Hence, we linked the words with the = symbol.

The number of link types after the first round of cross-lingual annotation (eng-jpn, eng-cmn) is summarized in Table 4. In the English-Japanese and English-Chinese corpora, 51.38% and 60.07% of the concepts have the same synsets: that is, slightly over half of the concepts can be directly translated. Around 5% of the concepts in the two corpora are linked to words close in the hierarchy (hyponym/hypernym). There were very few antonyms (0.5%). Similar relations plus translation equivalents account for 42.85% and 34.74% in the two corpora respectively. These parts are the most challenging for machine translation.

In this first round, when the annotator attempted to link concepts, it was sometimes the case that the translation equivalent was a word not excluded from wordnet by design. Especially common was cases of common nouns in Japanese and Chinese being linked to pronouns in English. In studying how concepts differ across languages, we consider these of interest. We therefore expanded our tagging effort to include pronouns.

6 Discussion and Future Work

The pilot study showed clearly that cross-lingual annotation was beneficial not just in finding interesting correspondences across languages but also in improving the monolingual annotation. In particular, we found many instances of multiword expressions that had been missed in the monolingual annotation. Using a wordnet to sense tag a corpus is extremely effective in improving the quality of the wordnet, and tagging and linking parallel text

Type	Eng-Jpn		Eng-Cmn	
linked	2,542		2,535	
=	1,416	51.58	1,712	60.07
∼	990	36.07	862	30.25
≈	186	6.78	128	4.49
⊃	75	2.73	94	3.30
$\supset^2$	8	0.81	13	1.51
⊂	63	2.30	39	1.37
$\subset^2$	10	1.01	18	2.09
!	1	0.04	2	0.07
#	14	0.51	13	0.46
unlinked	2,583		1,898	

Table 4: Analysis of links

is an excellent way to improve the quality of the monolingual annotation. Given how many problems we found in both wordnet and corpus when we went over the bilingual annotation, we hypothesize that perhaps one of the reasons WSD is currently so difficult is that the gold standards are not yet fully mature. They have definitely not yet gone through the series of revisions that many syntactic corpora have, even though the tagging scheme is far harder.

For this project, we improved our annotation process in two major ways:

(i) We expanded the scope of the annotation to include pronouns and named entities interrogatives. These will now be tagged from the monolingual annotation stage.

(ii) We improved the tool to make it possible to add new entries directly to the wordnets, so that they are available for tagging the remaining text. Using the comments to add new sense was a bad idea: synset-ids were cut and pasted, often with a character missing, and annotators often mistyped the link type. In addition, for words that appeared many times, it was tedious to redo it for each word. We are now testing an improved interface where annotators add new words to the wordnet directly, and these then become available for tagging. As a quality check, the new entries are reviewed by an expert at the end of each day, who has the option of amending the entry (and possibly re-tagging).

We are currently tagging the remaining texts shown in Table 1, with a preliminary release scheduled for September 2013. For this we are also investigating ways of improving the automatic cross-lingual annotation: using word level alignments; using global translation models and by relaxing the mapping criteria (in particular allowing linking across parts of speech through derivational links). When we have finished, we will also link the Japanese to the Chinese, using English as a pivot. Finally, we will go through the non-aligned concepts, and analyze why they cannot be aligned.

In future work we intend to also add structural semantic annotation to cover issues such as quantification. Currently we are experimenting with Dependency Minimal Recursion Semantics (DMRS: Copestake et al., 2005; Copestake, 2009) and looking at ways to also constrain these cross-linguistically (Frermann and Bond, 2012).

An interesting further extension would be to look at a level of discourse marking. This would be motivated by those translations which cannot be linked at a lower level. In this way we would become closer to the Groningen Meaning Bank, which annotates POS, senses, NE, thematic roles, syntax, semantics and discourse (Basile et al., 2012).

7 Conclusions

This paper presents preliminary results from an ongoing project to construct large-scale sense-tagged parallel corpora. Four languages are chosen for the corpora: English, Chinese, Japanese, and Indonesia. The annotation scheme is divided into two phrases: monolingual sense annotation and multilingual concept alignment. A pilot study was carried out in Chinese, English and Japanese for the short story *The Adventure of the Dancing Men*. The results show that in the English-Japanese and English-Chinese corpora, over half of the concepts have the same synsets and thus can be easily translated. However, 42.85% and 34.74% of the concepts in the two corpora cannot be directly linked, which suggests it is hard for machine translation. All annotated corpora will be made freely available through the NTU-MC, 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*.

References

Valerio Basile, Johan Bos, Kilian Evang, and Noortje Venhuizen. 2012. Developing a large semantically annotated corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, pages 3196–3200. Istanbul, Turkey.

Luisa Bentivogli, Pamela Forner, and Emanuele Pianta. 2004. Evaluating cross-language annotation transfer in the MultiSemCor corpus. In *20th International Conference on Computational Linguistics: COLING-2004*, pages 364–370. Geneva.

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 Ryan Foster. 2013. Linking and extending an open multilingual wordnet. In *51st Annual Meeting of the Association for Computational Linguistics: ACL-2013*. Sofia.

Francis Bond, Sanae Fujita, and Takaaki Tanaka. 2006. The Hinoki syntactic and semantic treebank of Japanese. *Language Resources and Evaluation*, 40(3–4):253–261. URL `http://dx.doi.org/10.1007/s10579-007-9036-6`, (Special issue on Asian language technology; re-issued as DOI s10579-008-9062-z due to Springer losing the Japanese text).

Francis Bond, Sanae Fujita, and Takaaki Tanaka. 2008. The Hinoki syntactic and semantic treebank of Japanese. *Language Resources and Evaluation*, 42(2):243–251. URL `http://dx.doi.org/10.1007/s10579-008-9062-z`, (Re-issue of DOI 10.1007/s10579-007-9036-6 as Springer lost the Japanese text).

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.

Arthur Conan Doyle. 1905. *The Return of Sherlock Homes*. George Newnes, London. Project Gutenberg `www.gutenberg.org/files/108/108-h/108-h.htm`.

Ann Copestake. 2009. Slacker semantics: Why superficiality, dependency and avoidance of commitment can be the right way to go. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 1–9. Athens.

Ann Copestake, Dan Flickinger, Carl Pollard, and Ivan A. Sag. 2005. Minimal Recursion Semantics. An introduction. *Research on Language and Computation*, 3(4):281–332.

Oliver Čulo, Silvia Hansen-Schirra, Stella Neumann, and Mihaela Vela. 2008. Empirical studies on language contrast using the English-German comparable and parallel CroCo corpus. In *Proceedings of Building and Using Comparable Corpora, LREC 2008 Workshop, Marrakesh, Morocco*, volume 31, pages 47–51.

Lea Cyrus. 2006. Building a resource for studying translation shifts. In *Proceedings of The Second International Conference on Language Resources and Evaluation (LREC-2006)*.

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

Lea Frermann and Francis Bond. 2012. Cross-lingual parse disambiguation based on semantic correspondence. In *50th Annual Meeting of the Association for Computational Linguistics: ACL-2012*, pages 125–129. Jeju, Korea.

Hitoshi Isahara, Francis Bond, Kiyotaka Uchimoto, Masao Utiyama, and Kyoko Kanzaki. 2008. Development of the Japanese WordNet. In *Sixth International conference on Language Resources and Evaluation (LREC 2008)*. Marrakech.

Kaname Kasahara, Hiroshi Sato, Francis Bond, Takaaki Tanaka, Sanae Fujita, Tomoko Kanasugi, and Shigeaki Amano. 2004. Construction of a Japanese semantic lexicon: Lexeed. In *IPSG SIG: 2004-NLC-159*, pages 75–82. Tokyo. (in Japanese).

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT Summit X*.

Sadao Kurohashi and Makoto Nagao. 2003. Building a Japanese parsed corpus — while improving the parsing system. In Anne Abeillé, editor, *Treebanks: Building and Using Parsed Corpora*, chapter 14, pages 249–260. Kluwer Academic Publishers.

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

Mingqin Li, Juanzi Li, Zhendong Dong, Zuoying Wang, and Dajin Lu. 2003. Building a large Chinese corpus annotated with semantic dependency. In *Proceedings of the second SIGHAN workshop on Chinese language processing-Volume 17*, pages 84–91. Association for Computational Linguistics.

George Miller. 1998. Foreword. In Fellbaum (1998), pages xv–xxii.

Nurril Hirfana Mohamed Noor, Suerya Sapuan, and Francis Bond. 2011. Creating the open Wordnet Bahasa. In *Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation (PACLIC 25)*, pages 258–267. Singapore.

Hwee Tou Ng and Hian Beng Lee. 1996. Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics*, pages 40–47.

Lars Nygaard and Jörg Tiedemann. 2003. OPUS — an open source parallel corpus. In *Proceedings of the 13th Nordic Conference on Computational Linguistics*.

Sebastian Padó and Katrin Erk. 2010. Translation shifts and frame-semantic mismatches: A corpus analysis. Ms: `http://www.nlpado.de/~sebastian/pub/papers/ijcl10_pado_preprint.pdf`.

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

Singapore Tourist Board. 2012. Your Singapore. Online: `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.

Jörg Tiedemann. 2009. News from OPUS — a collection of multilingual parallel corpora with tools and interfaces. In N. Nicolov, K. Bontcheva, G. Angelova, and R. Mitkov, editors, *Recent Advances in Natural Language Processing*, volume 5, pages 237–248. John Benjamins, Amsterdam/Philadelphia.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218.

Jörg Tiedemann and Lars Nygaard. 2004. The OPUS corpus — parallel and free. In *In Proceeding of the 4th International Conference on Language Resources and Evaluation (LREC-4)*.

Martin Volk, Anne Göhring, Torsten Marek, and Yvonne Samuelsson. 2010. SMULTRON (version 3.0) — The Stockholm MULtilingual parallel TReebank. `http://www.cl.uzh.ch/research/paralleltreebanks_en.html`.

Gan Kok Wee and Tham Wai Mun. 1999. General knowledge annotation based on how-net. *Computational Linguistics and Chinese Language Processing*, 4(2):39–86.

Yunfang Wu, Peng Jin, Yangsen Zhang, and Shiwen Yu. 2006. A chinese corpus with word sense annotation. In *Computer Processing of Oriental Languages. Beyond the Orient: The Research Challenges Ahead*, pages 414–421. Springer.

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.