Building The Sense-Tagged Multilingual Parallel Corpus

Shan Wang, Francis Bond

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

Abstract

Sense-annotated parallel corpora play a crucial role in natural language processing. This paper introduces our progress in creating such a corpus for Asian languages using English as a pivot, which is the first such corpus for these languages (Chinese, Japanese and Indonesian). Two sets of tools have been developed for sequential and targeted tagging, which are also easy to be set up for any new languages. This paper also briefly presents the general guidelines for doing this project. The current results of the monolingual sense-tagging and multilingual linking are illustrated, which indicate the differences among genres and language pairs. All the tools, guidelines and the manually annotated corpus will be freely available at http://compling.ntu.edu.sg/ntumc.

Keywords: sense-tagging, multilingual corpus, parallel corpus

1. Introduction

Semantically annotated corpora are of significant values in natural language processing. In particular, sense annotated corpora based on Princeton Wordnet (Fellbaum 1998) have been widely developed (Petrolito & Bond 2014). One such corpus is English SemCor (Landes et al. 1998), which is among the early sense-tagged corpora. After it was created, Italian, Romanian and Japanese translations of it have been made and sense-tagged (Bentivogli & Pianta 2005; Lupu et al. 2005; Tan & Bond 2012). Such kind of Semcors have been used in a large number of tasks (Kilgarriff 1998; Gonzalo et al. 2000; Navigli et al. 2003; Gutiérrez et al. 2011). However, there is no such resource for Asian languages.

Instead of translating the English SemCor to Asian languages, we made use of the Nanyang Technological University Multilingual Corpus (NTU-MC) which contains 595,000 words (26,000 sentences) in seven languages (Arabic, Chinese, English, Indonesian, Japanese, Korean and Vietnamese) from seven language families (Afro-Asiatic, Sino-Tibetan, Indo-European, Austronesian, Japonic, Korean as a language isolate and Austro-Asiatic) (Tan & Bond 2012; Bond et al. 2013). We selected four of these languages for further annotation: English, Chinese, Japanese, and Indonesian. The corpus of each language was first manually sense tagged with Princeton Wordnet (Fellbaum 1998), Chinese Open Wordnet (Wang & Bond 2013a, 2013b), Japanese Wordnet (Isahara et al. 2008) and Wordnet Bahasa (Nurril Hirfana et al. 2011), and then linked to the English corpus at the concept level respectively (Bond et al. 2013; Bond & Wang 2014). To the best of our knowledge, this is the first such multilingual corpus for these Asian languages. All the tools, guidelines and annotated corpus will be freely available at http://compling.ntu.edu.sg/ntumc. By doing this project, we aim to provide a useful resource for the community.

The following sections are arranged as follows. Section 2 introduces the tools, guidelines and quality control of the corpus. The current results of the annotated corpus are illustrated in Section 3. Section 4 summarizes this paper and gives directions for future work.

2. Building Sense-tagged Multilingual Parallel Corpora

Though there are some parallel corpora (Koehn 2005; Cyrus 2006; Čulo et al. 2008; Volk et al. 2010) and sense-tagged corpora (Ng & Lee 1996; Mingqin et al. 2003), multilingual sense-tagged corpora are rare. The only one we know of is English SemCor and its translations into Italian, Romanian and Japanese. This project aims for creating a sense-tagged parallel corpus for Asian languages by utilizing the texts of NTU-MC. The current size of the corpus we are tagging is shown in Table 1. There are 7,093 sentences in the English texts, which are translated into Chinese, Japanese and Indonesian, making a total of 22,762 sentences. Words are all the tokens, while concepts refer to content words and multiword expressions (MWE). The actual number is changing as the project goes on.

With this project going on, we are aware of the respects which can speed up the development of such tasks: (i) convenient annotation tools, (ii) clear and detailed guidelines, (iii) follow-up checking to guarantee quality control. All data are manually annotated by trained linguistic students.

2.1 Annotation Tools

We developed two sets of annotation tools: one for sequential/textual tagging (sentence by sentence) and one for targeted/lexical tagging (word by word) (Langone et al. 2004). The former is illustrated in Figure 1, which embeds

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 (yoursing)	2,988	2,332	2,723	2,197	74,300	32,600
Total		7,093	6,438	7,034	2,197	169,800	74,600

Table 1. NTU-MC Size

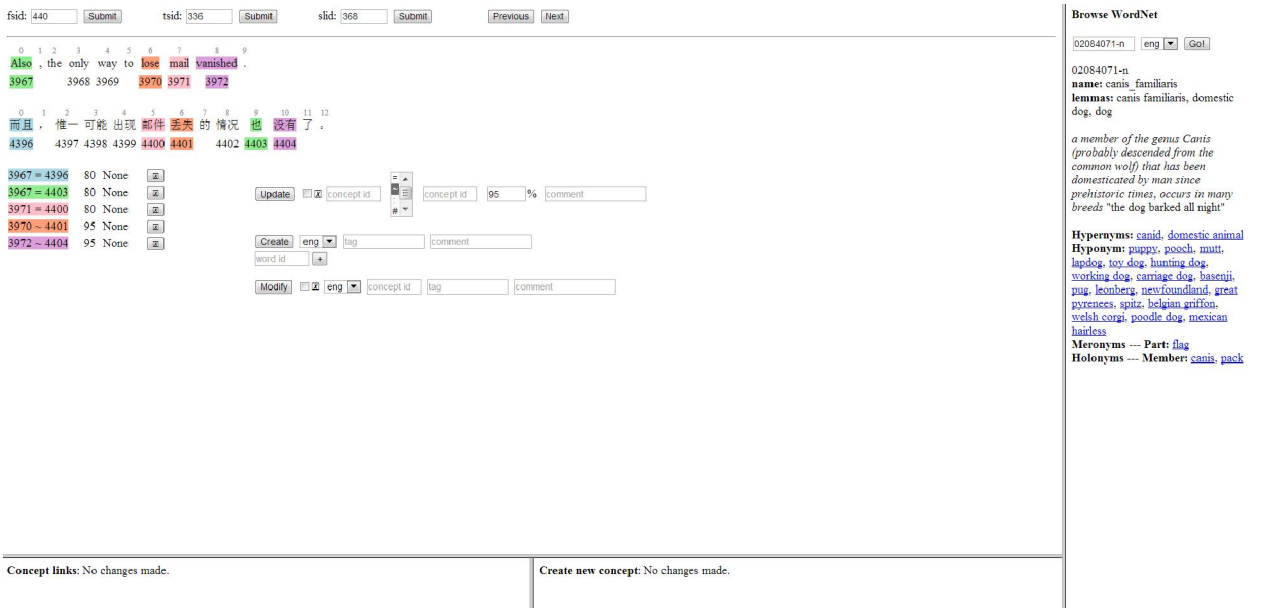

Figure 1. The sequential tagging tool

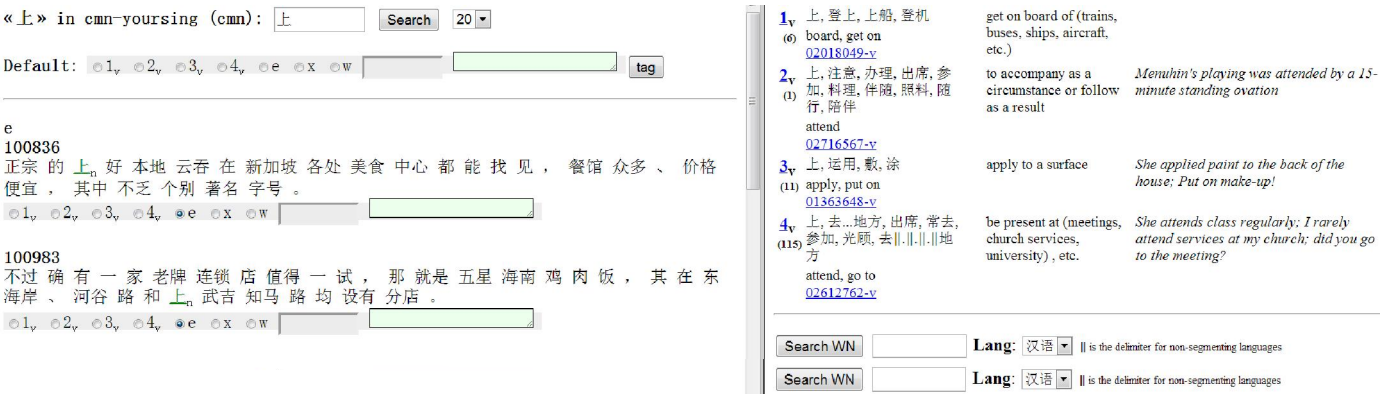

Figure 2. The lexical tagging tool

two functions: sense tagging and concept linking. On the right-hand side, annotators can choose an existing wordnet sense. On the left side, they can link the concepts and then all the linked concepts are automatically shown in color.

The lexical tagging tool is illustrated in Figure 2. The left side shows the sentences with the target words to be tagged. The right side shows the lemmas in Chinese Open Wordnet (Wang & Bond 2013a, 2013b). If none of them is suitable, annotators can search wordnets through choosing languages, such as English, Japanese, and Indonesian.

2.2 General Guidelines

Because this project is comprised of multiple tasks, we formulated our guidelines accordingly. We experienced the process of making-using-revising-using again in the formation of our guidelines.

Taking the English-Chinese corpus part as an example, this section introduces the general guidelines. **First, pre-processing guidelines.** Three tasks are carried out during pre-processing: word segmentation (WS), Part-of-Speech (POS) tagging and concept identification.

WS and POS tagging are critical pre-requisite steps for numerous Chinese information processing tasks, such as parsing, machine translation and information extraction. Various methods have been proposed for WS using supervised methods (Xue 2003) or unsupervised methods (Sun et al. 1998; Wang et al. 2008). We used Stanford Chinese Segmenter and POS Tagger[1] to do the two tasks based on the CTB standards, which was designed to build parse trees. Since different applications need different WS systems (Song 1997), some amendments were carried out to CTB standards.

(I) Word Segmentation

The aim of doing the pre-processing is to facilitate sense tagging and machine translation. For WS, we obey these general criteria:

(i) Compositional or not

If a unit is non-compositional, it is regarded as a word, such as 黑板 *hēibǎn* black-board 'blackboard', 白菜 *báicài* white-vegetable 'cabbage'. Some units have multiple meanings, with one as non-compositional and the other as compositional. 爱人 *àirén* is a case in point, when it means "spouse", it is non-compositional and thus is taken as a word; when it means "love people", it is compositional and thus is taken as two words 爱/人 *ài/rén*.

(ii) Productivity

If a unit has unlimited productivity, it is regarded as more than one word. For example, 买/书 *mǎi/shū* 'buy book (s)', 买/饭 *mǎi/fàn* 'buy food', 买/水 *mǎi/shuǐ* 'buy water'. On the contrary, some units have limited productivity, so we treat them as words, such as 12 months in a year, Chinese zodiac (鼠年 *shǔnián* 'year of the Rat', 蛇年 *shénián* 'year of the Snake', etc.), and first 10 days in a month of the Chinese calendar (初一 *chūyī* 'first day of a month', 初二 *chūèr* 'second day of a month', etc.)

(iii) Containing a bound morpheme or not

If a unit contains a bound morpheme, it is regarded as a WS unit. [2] For example, 桌上 *zhuōshàng* 'on the table' is one segmentation unit, while 桌子/上 *zhuōzi/shàng* 'on the table' is two segmentation units.

(vi) Affixes

If a word has an affix, we do not segment the affix[3]. For example, words with a prefix: *超*大 *chāodà* 'super large', *超*强 *chāoqiáng* 'super powerful', *超*好 *chāohǎo* 'super good'; words with a suffix: 现代*化* *xiàndàihuà* 'modernize', 国际*化* *guójìhuà* 'internationalize'.

(v) Frequency

If a unit is highly frequent, it will be considered as one WS unit. For example, 空中 *kōngzhōng* air-middle 'in the air', 听见 *tīngjiàn* listen-see 'hear'.

(II) POS Tagging

Regarding POS tagging, because we are using wordnet senses to tag our corpus, we only need four tags: noun, verb, adjective, adverb. Thus we made some amendments to the CTB tagsets of content words, as shown in Table 2[4]. For example, "JJ" in CTB refers to noun-modifier other than nouns, which makes this class divergent, including adjectives and verbs. To do the sense-tagging, we must have more fine-grained POS than just "JJ". Please note that the comparison in Table 2 assumes that CTB tags are correctly assigned to words. For example, CDs are natural numbers used to measure the size of sets.[5] The tagger wrongly tags 多 *duō* 'many' as CD, which should be an adjective. Therefore, the comparison does not say that in our system CD should be adjectives for cases like 多 *duō* 'many'.

There are also some POS in CTB that we do not tag: (i) functional words: AS, CC, CD, CS, DEC, DEG, DER, DEV, MSP, P, PU, SP; (ii) BA (把 *bǎ* in ba-construction), LB (被 *bèi* in long bei-construction) and SB (被 *bèi* in short bei-construction): some research treat 把 *bǎ* and 被 *bèi* as prepositions and some as verbs; we have not decided how to deal with them, so currently we do not tag them; (iii) IJ (interjections): they do not contribute much to the semantic system, so we do not tag them.

(III) Concept Identification

The concepts are the basic units in doing annotation, so we must identify them, including content words and WEs.

[1] http://nlp.stanford.edu/software/segmenter.shtml

[2] Here we use "word segmentation unit" in order not to fall into the trouble of deciding whether it is a word or not, which is a very controversial issue in Chinese. This is common practice in Chinese language processing.

[3] This may tend to change depending on the whether the affixes are productive in forming words.

[4] Not all examples are from NTU-MC.

[5] http://en.wikipedia.org/wiki/Cardinal_number

Tagset	English Explanation	In Our System	Example
AD	adverb	adverb	也许 *yěxǔ* 'perhaps'
CD	cardinal number	noun	四 个 小时 *sì gè xiǎoshí* 'four hours'
DT	determiner	noun	这 *zhè* 'this'
ETC	tag for words 等 *děng* 'etc.', 等等 *děngděng* 'etc.' in coordination phrase	adverb	等 *děng* 'etc.'
FW	foreign words	It depends.	popiah: noun
JJ	noun-modifier other than nouns	adjective	黑 咖啡 *hēi kāfēi* 'black coffee'
		verb	专有 作品 *zhuān yǒu zuòpǐn* exclusively-have works 'Proprietary works'
LC	localizer	noun	上 *shàng* 'on'
M	measure word (including classifiers)	noun	杯 *bēi* 'cup'
NN	common noun	noun	蛋糕 *dàn'gāo* 'cake'
NR	proper noun	noun	新加坡 *Xīnjiāpō* 'Singapore'
NT	temporal noun	noun	今天 *jīntiān* 'today'
OD	ordinal number	adjective	第一 站 *dìyī zhàn* 'first station'
		noun	获得 第一 *huòdé dìyī* 'get the first'
		adverb	第一， 属于 国内 首创 *dìyī, shǔyú guónèi shǒuchuàng* 'First, belong to domestic innovation'
ON	onomatopoeia	adverb	哗哗 地 流 *huāhuā de liú* gurglingly de flow 'flow gurglingly'
		adjective	睡 得 呼呼 的 *shuì de hūhū de* sleep de snore "catch some Z's"
		verb	冰箱 嗡 了 一下。*Bīngxiāng wēng le yīxià* 'The refrigerator hummed once.'
PN	pronoun	noun	他们 *tāmen* 'they'
VA	predicative adjective	adjective	欢乐 *huānlè* 'joyous'
VC	copula 是 *shì*	verb	是 *shì* 'be'
VE	有 *yǒu* as the main verb	verb	有 *yǒu* 'have'
VV	other verbs	verb	购买 *gòumǎi* 'purcase'

Table 2. Comparison between CTB tagsets and our tags

This makes the corpus have the characteristic of multi-layer annotation.

Second, sense tagging guidelines. Annotators are required to choose an existing wordnet sense or one of the 5 taggers: (i) p POS that should not be tagged; (ii) e error in tokenization; (iii) s missing sense (not in wordnet); (iv) u lemma not in wordnet but POS open class (tagged automatically); (v) m Multiword: if the lemma is a multiword, this tag means it is not appropriate; if the lemma is single-word, this tag means it should be part of a multiword.

After a round of tagging, annotators reported their confusion with the taggers s, u, m, p, so we merged s and u to w, meaning it should be in wordnet, but is not; m and p to x, meaning it does not need to be tagged. The agreement between annotators in using simplified tags when tagging the short story "The Adventure of the Speckled Band" is shown in Table 3. A, B are from a semantic class this semester using the simplified tags. C is the data annotated by a linguistic undergraduate using the older tags. S is the silver standard (majority of A, B, C).

Agreement	Agreement rate (%)
A vs S	0.720
B vs S	0.718
C vs S	0.612
A vs B	0.598

Table 3. The agreement rate between annotators

Third, parallel corpus guidelines. Since the senses of the concept have been tagged before doing the cross lingual linking, we automatically linked the concepts in Chinese, Japanese and Indonesian corpus to the concepts in the English corpus respectively when two concepts in the paralleled sentences have the same synsets. Other than this, annotators are required to choose from one of these

symbols: (i) = same synset, (ii) < hyponym, (iii) > hypernym, (iv) ~ lexically linked, (v) ≈ pragmatically linked, (vi) ! antonym, (vii) # weak antonym. Meanwhile, they can revise the wordnet sense annotation as they do cross lingual linking.

Besides these symbols, it is also important to note what kinds of elements should be linked. One of such cases is whether to link bare nouns or the determiner "the"+noun. It is common that Chinese and Japanese use bare nouns while English need either a plural form or the determiner "the" before the noun. For example: (i) English uses a plural form while Chinese is in bare form: Tigers are striped. 老虎有斑纹。*Lǎohǔ yǒu bānwén.* (ii) English uses a determiner while Chinese is in bare form: The table is white. 桌子是白的。*Zhuōzi shì bái de.*

In creating the parallel corpus, the quantification is not quite relevant, so our solution to this question is not linking the determiners, except in the very special cases of names with "the" in them, like "The Hague" and "The the". Therefore, in (1), which is from the essay, we only link "planet" (not "the planet") and 地球 *chikyu* 'earth'.

(1) a. …… all over the **planet** …… (essay)
b. …… **地球** 全体 に……
chikyu zentai ni
earth whole in

The other issue is to link at the word level or MWE level.

(2) a. ……when this fellow **comes again** …… (Dancing Men)
b. 等 那 个 家伙 **再来**
děng nà gè jiāhuo zàilái
wait for that CL guy come again

In (2), 再来 *zàilái* 'come again' is a commonly used unit in Chinese, while "come again" is a productive construction "v+adv" in English. In such a case, we can link 再 *zài* =again, 来 *lái* = come, or 再来 *zàilái* = come again, making the English an MWE. Our solution to this issue depends on the WS result. If 再来 *zàilái* is taken as one word, we only link it to "come again", and not doing the individual word linking.

A third issue is whether to link some words with their aspectual markers (着 *zhe* 'progressive aspect', 了 *le* 'perfective aspect', 过 *guo* 'experiential aspect') in Chinese.

(3) a. Finally he **led** the way into the drawing-room …… (Dancing Men)
b. 最后 他 **领 着** 我们 去 客厅
zuìhòu tā lǐng zhe wǒmen qù kètīng
finally he lead ASP us go to drawing-room

(4) a. …… **across** the paper upon which they are drawn. (Dancing Men)
b. 在 纸 上 **横 着** 画 了……
zài zhǐ shàng héng zhe huà le
on paper on horizontal ASP draw ASP

In (3) 领着 *lǐng zhe* 'lead-ASP' is the verb 领 *lǐng* 'lead' followed by the aspectual marker 着 *zhe* 'progressive aspect' to indicate the action is in process. In (4), 横着 *héng zhe* 'horizontal-ASP' is the adjective 横 *héng* 'horizontal' followed by 着 *zhe* 'progressive aspect' to indicate the state. In (3)b, it is fine only to use 领 *lǐng* 'lead' in the Chinese sentence, while in (4)b, in order to modify the verb 画 *huà* 'draw', it must be 横着 *héng zhe* 'horizontal-ASP' rather than 横 *héng* 'horizontal'.

Aspectual makers are extremely frequent words in Chinese, so in (3) and (4), we only link 领 *lǐng* to "lead" and 横 *héng* to "across".

A fourth issue is whether to create an equal link or ~ link.

(5) a. I am fairly familiar with **all** forms of secret writings …… (Dancing Men)
b. 我 比较 熟悉 **各 种**
Wǒ bǐjiào shúxī gè zhǒng
I comparatively be familiar with every kind

形式 的 秘密 文字……
xíngshì de mìmì wénzì
form DE secret writing

In (5)b, 各种 *gè zhǒng* 'every kind' is a commonly used unit. 各 *gè* means "every", 种 *zhǒng* means "kind". It is fine to link 各 *gè* ~ all, or 各种 *gè zhǒng* ~ all. For such a case, we will still depend on the WS result. If 各种 *gè zhǒng* is segmented, we link 各 *gè* ~all; if it is treated as one unit, we link 各种 *gè zhǒng* ~ all.

2.3 Quality Control

After the first round of annotation, in order to guarantee the quality, we are now carrying out follow-up checking using the lexical tagging method. In this round, we are mainly concerned about these issues: inconsistency, errors, concepts not sense-tagged, concepts unlinked. Thus we developed another set of tools for sense-tagging and multilingual linking as illustrated in Figure 2.

3. Current Results of the Project

Sense-tagging is a crucial part of this project. Table 4 shows the percentage of the monolingually tagged senses. Up to now, most concepts that exist in wordnets have been tagged. Missing senses or words in wordnets hinder the annotation, so wordnets with high accuracy and coverage would speed up the process. Our corpus thus in turn is a good source to improve wordnets through providing data that can be added to each language's wordnet.

Genre	English	Chinese	Japanese	Indonesian
Essay	82.5	77.2	82.3	---
Story	84.4	69.2	71.9	---
Tourism	80.8	73.2	---	75.7

Table 4. Percentage (%) of tagged senses

Link	Story (Eng-Cmn)		Story (Eng-Jpn)		Essay (Eng-Cmn)		Essay (Eng-Jpn)		Tourism (Eng-Ind)	
	No.	percent (%)	No.	percent (%)	No.	percent (%)	No.	percent (%)	No.	percent (%)
=	2,765	43.6	2,632	48.7	1,401	31.7	2,203	34.4	14,156	60
<	122	1.9	96	1.8	1	0	79	1.2	230	1
>	267	4.2	151	2.8	0	0	52	0.8	330	1.4
~	2,024	31.9	2,045	37.8	3,010	68.1	2,830	44.2	5,618	23.8
≈	1,144	18	454	8.4	0	0	1,178	18.4	3,215	13.6
!	13	0.2	2	0	11	0.2	17	0.3	24	0.1
#	13	0.2	23	0.4	0	0	42	0.7	12	0.1
Total	6,348	100	5,403	100	4,423	100	6,401	100	23,585	100

Table 5. Number of Links

Concept level parallel is very important for machine translation. Table 5 illustrates all the linked concepts in three genres of five language pairs. Out of the seven types of links in each genre, *story* and *tourism* data have most = links, while *essay* has more ~ links. This means that the essay tends to get a freer translation compared to stories and tourism. Those linked with ≈ and # are the most difficult for machine translation, because they are very weak connections.

4. Conclusions and Future Work

This paper presents our progress in sense-tagging and linking a multilingual corpus, which is the first such corpus for multiple Asian languages. During the process, two sets of tools are developed for sequential and targeted tagging respectively. These tools are easy to set up for any new languages. This paper also introduces the general guidelines for doing this project. We have illustrated the current results of monolingual sense-tagging and multilingual linking, which show the difference among genres and language pairs. All the tools, detailed guidelines and the manually annotated corpus will be freely available at http://compling.ntu.edu.sg/ntumc.

In future work, we will continue doing the checkup for the second round of annotation in order to guarantee the quality of the data. Meanwhile, we are adding new entries and new senses to the wordnets of these languages so as to improve these wordnets toward better coverage and accuracy. Furthermore, we will conduct a cross lingual study for the corpus and utilize them in NLP tasks to test their performance.

Acknowledgements

This research was supported by the MOE Tier 1 grant *Shifted in Translation—An Empirical Study of Meaning Change across Languages* (RG51/12) and the NTU HASS Incentive Scheme *Equivalent but Different: How Languages Represent Meaning in Different Ways*.

References

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

Bond, Francis, Timothy Baldwin, Richard Fothergill & Kiyotaka Uchimoto. 2012. *Japanese SemCor: A sense-tagged corpus of Japanese*. Proceedings of the 6th Global WordNet Conference. Matsue, Japan. *pp.56–63*.

Bond, Francis & Shan Wang. 2014. *Issues in building English-Chinese parallel corpora with wordnets*. Proceedings of The Seventh Global WordNet Conference (GWC-7), ed. by Heili Orav, Christiane Fellbaum & Piek Vossen. Tartu, Estonia. *pp.391-399*.

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

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

Cyrus, Lea. 2006. *Building a resource for studying translation shifts*. Proceedings of The Second International Conference on Language Resources and Evaluation (LREC-2). *pp.1240-1245*.

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

Gonzalo, Julio, Irina Chugur & Felisa Verdejo. 2000. *Sense clusters for information retrieval: evidence from Semcor and the EuroWordNet InterLingual index*. Proceedings of the ACL-2000 workshop on Word senses and multi-linguality: Association for Computational Linguistics. *pp.10-18*.

Gutiérrez, Yoan, Sonia Vázquez & Andrés Montoyo. 2011. *Improving WSD using ISR-WN with relevant semantic trees and SemCor senses frequency*. Proceedings of Recent Advances in Natural Language Processing. Hissar, Bulgaria. *pp.233-239*.

Isahara, Hitoshi, Francis Bond, Kiyotaka Uchimoto, Masao Utiyama & Kyoko Kanzaki. 2008. *Development of the Japanese WordNet*. Proceedings of the Sixth International Conference on Language Resources and Evaluation, ed. by Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis & Daniel Tapias. Marrakech: European Language Resources Association. *pp.2420-2423*.

Kilgarriff, Adam. 1998. *Senseval: An exercise in evaluating word sense disambiguation programs*. Proceedings of the First International Conference on Language Resources and Evaluation: Citeseer. *pp.581-588*.

Koehn, Philipp. 2005. *Europarl: A parallel corpus for statistical machine translation*. Conference Proceedings: The tenth Machine Translation Summit. Phuket, Thailand. *pp.79-86*.

Landes, Shari, Claudia Leacock & Christiane Fellbaum. 1998. *Building semantic concordances*. WordNet: An Electronic Lexical Database, ed. by Christiane Fellbaum: MIT Press. *pp.199–216*.

Langone, Helen, Benjamin R Haskell & George A Miller. 2004. *Annotating wordnet*. Proceedings of Frontiers in Corpus Annotation Workshop, Workshop at HLT-NAACL 2004.

Lupu, Monica, Diana Trandabat & Maria Husarciuc. 2005. *A Romanian SemCor aligned to the English and Italian MultiSemCor*. Proceedings of 1st ROMANCE FrameNet Workshop at EUROLAN 2005 Summer School. EuroLAN, Cluj-Napoca, Romania. *pp.20-27*.

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

Navigli, Roberto, Paola Velardi & Aldo Gangemi. 2003. Ontology learning and its application to automated terminology translation. *Intelligent Systems, IEEE* 18 (1). *pp.22-31*.

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

Nurril Hirfana, Bte Mohamed Noor, Sapuan Suerya & Francis Bond. 2011. *Creating the Open Wordnet Bahasa*. Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation, ed. by Helena Hong Gao & Minghui Dong. Singapore. *pp.255-264*.

Petrolito, Tommaso & Francis Bond. 2014. *A survey of wordnet annotated corpora*. Proceedings of The Seventh Global WordNet Conference (GWC-7), ed. by Heili Orav, Christiane Fellbaum & Piek Vossen. Tartu, Estonia. *pp.236-245*.

Song, Rou. 1997. Exploring word segmentation standards (关于分词规范的探讨). *Applied Linguistics (语言文字应用)* (3). *pp.111-112*.

Sun, Maosong, Dayang Shen & Benjamin K. Tsou. 1998. *Chinese word segmentation without using lexicon and hand-crafted training data*. Proceedings of the 17th international conference on Computational linguistics-Volume 2: Association for Computational Linguistics. *pp.1265-1271*.

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

Volk, Martin, A Göhring, Torsten Marek & Yvonne Samuelsson. 2010. SMULTRON (version 3.0)-The Stockholm MULtilingual parallel TReebank.

Wang, Shan & Francis Bond. 2013a. *Building the Chinese Open Wordnet (COW): starting from core synsets*. Proceedings of The 11th Workshop on Asian Language Resources, Workshop of The 6th International Joint Conference on Natural Language Processing (IJCNLP-6). Nagoya, Japan. *pp.10-18*.

Wang, Shan & Francis Bond. 2013b. Theoretical and practical issues in creating Chinese Open Wordnet (COW). *Paper presented at The 7th International Conference on Contemporary Chinese Grammar (ICCCG-7)*, Nanyang Technological University, Singapore.

Wang, Zhenxing, Changning Huang & Jingbo Zhu. 2008. *The character-based CRF segmenter of MSRA&NEU for the 4th Bakeoff*. Proceedings of The Sixth SIGHAN Workshop on Chinese Language Processing. *pp.98-101*.

Xue, Nianwen. 2003. Chinese word segmentation as character tagging. *Computational Linguistics and Chinese Language Processing* 8 (1). *pp.29-48*.