IMI — A Multilingual Semantic Annotation Environment

Francis Bond, Luís Morgado da Costa and Tuấn Anh Lê
Linguistics and Multilingual Studies
Nanyang Technological University, Singapore
{bond@ieee.org, luis.passos.morgado@gmail.com, tuananh.ke@gmail.com}

Abstract

Semantic annotated parallel corpora, though rare, play an increasingly important role in natural language processing. These corpora provide valuable data for computational tasks like sense-based machine translation and word sense disambiguation, but also to contrastive linguistics and translation studies. In this paper we present the ongoing development of a web-based corpus semantic annotation environment that uses the Open Multilingual Wordnet (Bond and Foster, 2013) as a sense inventory. The system includes interfaces to help coordinating the annotation project and a corpus browsing interface designed specifically to meet the needs of a semantically annotated corpus. The tool was designed to build the NTU-Multilingual Corpus (Tan and Bond, 2012). For the past six years, our tools have been tested and developed in parallel with the semantic annotation of a portion of this corpus in Chinese, English, Japanese and Indonesian. The annotation system is released under an open source license (MIT).

1 Introduction

Plain text parallel corpora are relatively widely available and widely used in NLP, such as machine translation system development (Koehn, 2005, e.g.,). In contrast, there are very few parallel sense tagged corpora due to the expense of tagging the corpora and creating the sense inventories in multiple languages. The one exception is the translations of English SemCor (Landes et al., 1998) for Italian (Bentivogli and Pianta, 2005), Romanian (Lupu et al., 2005) and Japanese (Bond et al., 2012). Even for this corpus, not all of the original English texts have been translated and tagged, and not all words are tagged in the translated text (typically only those with a corresponding English sense).

In this paper we present IMI, a web-based multilingual semantic annotation system designed for the task of sense annotation. The main goals of its design were to decrease the cost of production of these resources by optimizing the speed of tagging, and to facilitate the management of this kind of project. To accomplish this, we aimed at developing a simple and intuitive web-based system that allows parallel tagging by many users at a time, optimized for speed by requiring minimum input from the annotators.

We centered our development around the annotation of the NTU-Multilingual Corpus (NTU-MC: Tan and Bond, 2012). The NTU-MC is an open multilingual parallel corpus originally designed to include many layers of syntactic and semantic annotation. We selected a portion of this corpus based on 7,093 sentences of English, totaling 22,762 sentences of Chinese, Japanese and Indonesian parallel text. A series of undergraduate linguistics students were trained on the tool and annotated the corpus over several years. They also offered extensive qualitative and quantitative feedback on the usage of our system.

The remainder of this paper is arranged as follows. In Section 2 we introduce related work. Section 3 describes the main functionality of our system then we finish with Section 4, which summarizes and discusses our current and future work.

2 Related Work

In this section we introduce the corpus (NTU-MC), the sense inventory (OMW), and a brief overview of currently available tools.

Proceedings of ACL-IJCNLP 2015 System Demonstrations, pages 7–12,
Beijing, China, July 26-31, 2015.

2.1 The NTU-Multilingual Corpus (NTU-MC)

The NTU-MC (Tan and Bond, 2012) has data available for eight languages from seven language families (Arabic, Chinese, English, Indonesian, Japanese, Korean, Vietnamese and Thai), distributed across four domains (story, essay, news, and tourism). The corpus started off with monolingual part-of-speech (POS) annotation and cross-lingual linking of sentences. We are extending it to include monolingual sense annotation and cross-lingual word and concept alignments (Bond et al., 2013). Out of the available languages, Chinese, English, Japanese and Indonesian were chosen for further processing and annotation (due to the availability of lexical and human resources). As part of the annotation, we are also expanding the sense and concept inventory of the wordnets: Princeton Wordnet (PWN: Fellbaum, 1998), the Japanese Wordnet (Isahara et al., 2008), the Chinese Open Wordnet (Wang and Bond, 2013) and the Wordnet Bahasa (Nurril Hirfana *et al.* 2011) through the Open Multingual Wordnet (Bond and Foster, 2013).

2.2 The Open Multilingual Wordnet

The task of semantic annotating a corpus involves the manual (and often automated) disambiguation of words using lexical semantic resources – selecting, for each word, the best match in a pool of available concepts. Among this type of resources, the PWN has, perhaps, attained the greatest visibility. As a resource, a wordnet is simply a huge net of concepts, senses and definitions linked through many different types of relations. Because of popularity and confirmed utility, many projects have developed wordnets for different languages.

The Open Multilingual Wordent (OMW) (Bond and Foster, 2013) is an open source multilingual resource that combines many individual open-source wordnet projects, along with data extracted from Wiktionary and the Unicode Common Locale Data Repository. It contains over 2 million senses distributed over more than 150 languages, linked through PWN. Browsing can be done monolingual or multilingually, and it incorporates a full-fledged wordnet editing system which our system uses (OMWEdit: da Costa and Bond, 2015).

2.3 Other Available Systems

There are many text annotation tools available for research (e.g., Stenetorp et al., 2012). However, sense annotation has some features that differ from most common annotation tasks (such as NE or POS annotation). In particular, the number of tags, and the information associated with each tag is very large. Sense tagging for English using the PWN, for example, when unrestricted, defaults at over a hundred thousand possible tags to chose from: even constrained by the lemma, there may be over 40 tags and the set of tags will very from lemma to lemma.

There are only a few annotation tools designed specifically for sense annotation. We were able to find the following: the tools to tag the Hinoki Corpus (Bond et al., 2008), for Japanese, and the Sense Annotation Tool for the American National Corpus (SATANiC: Passonneau et al., 2009), for English. Both of these tools were developed to be used in a monolingual environment, and have not been released.

The only open source tool that we could find was Chooser (Koeva et al., 2008), a multi-task annotation tool that was used to tag the Bulgarian Sense Tagged Corpus (Koeva et al., 2006). This tool is open source, language independent and is capable of integrating a wordnet as a sense inventory. Unfortunately, it was not designed to be a web-service which means it is difficult to coordinate the work of multiple users.

3 System Overview and Architecture

Given the scenario of available systems, we decided we had enough motivation to start the development of a new Semantic Annotation Environment (IMI).

Because a large part of sense-tagging is adding new senses to the inventory, we integrated IMI with the existing tools for editing and displaying the Open Multilingual Wordnet. This integration was done mainly through the development of a single web-based environment, with a common login, and API communications between interfaces. We also designed a custom mode to display OMW results in a condensed way.Sharing a common login system allows our annotators to access the OMW wordnet editing mode (right hand of Figure 1) so that, when needed, annotators can add new senses and concepts to fit the data in the corpus.

Our system is written in Python and uses SQLite

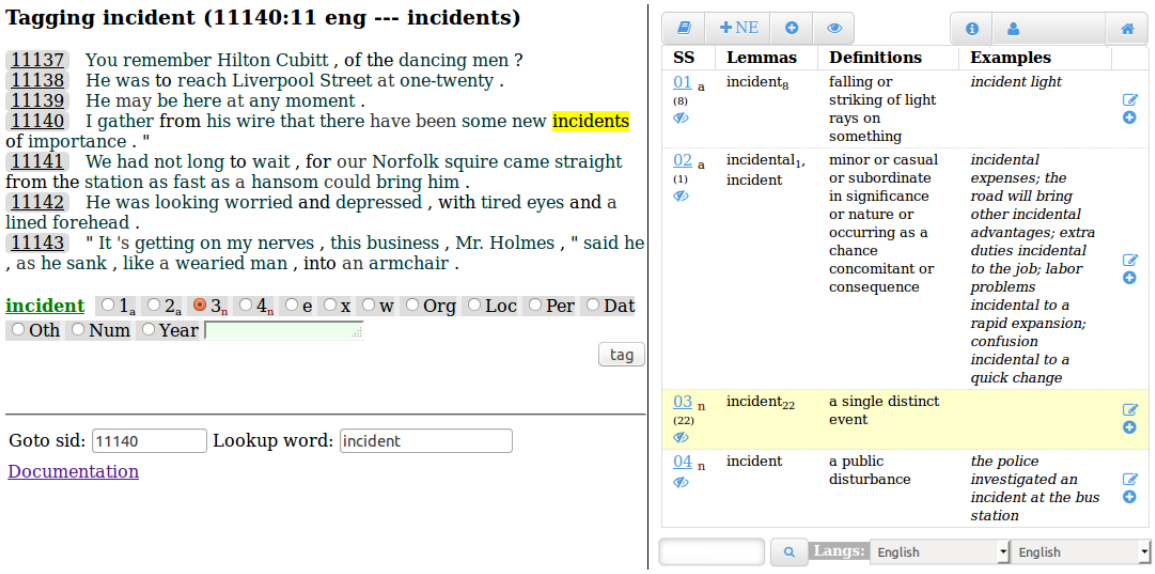

Figure 1: Sequential/Textual Tagger Interface

to store the data. It is tested on Firefox, Chrome and Safari browsers. In the remainder of this section we discuss its main functionality.[1]

3.1 The Annotation Interfaces

The sequential/textual tagger (Figure 1) was designed for concept-by-concept sequential tagging. It shows a short context around the sentence currently being tagged. Clicking a word generates an automated query in the OMW frame (on the right of Figure 1).

As it is costly to remember the set of senses for each word, we normally tag with a lexical/targeted tagger (Figure 2 displays only the left side of this tagging interface, as the OMW frame is identical to that of Figure 1). Querying the OMW with this tagger is very similar to the description above. The main difference of this interface is that it focuses on a single lexical unit across the corpus. In the example provided in Figure 2, every occurrence of the lemma ***wire*** is tagged at the same time. For frequent words, the number of results displayed can be restricted. In this interface, only the sentence where the word occurs is provided as context, but a larger context can also be accessed by clicking on the sentence ID. Since the concept inventory is the same for the full list of words to be tagged, time is saved by keeping the concepts fresh in the annotator's mind, and quality is ensured by comparing different usages of different senses at the same time.

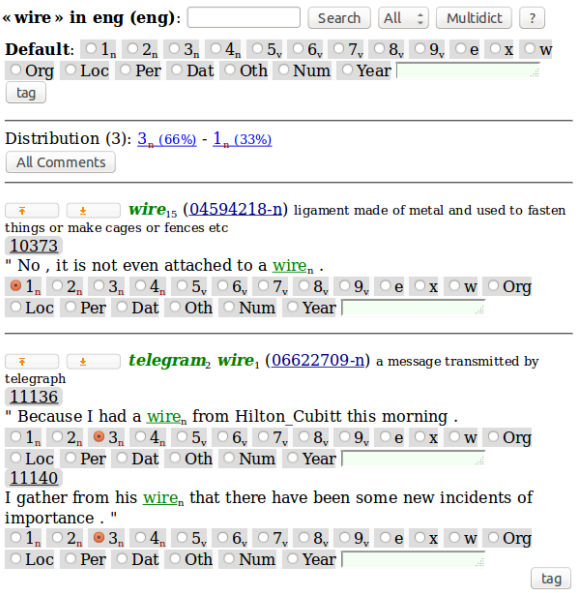

Figure 2: Targeted/Lexical Tagger

In both tagging interfaces, a tag is selected among an array of radio buttons displayed next to the words being tagged. Besides the numerical options that match the results retrieved by the OMW, the interface also allows tagging with a set of meta tags for named entities and to flag other issues. We use a similar set to that of Bond et al. (2013). With every tag, a comment field is provided as an optional field, where annotators can leave notes or describe errors.

[1] The annotation interface software and corpora are available from the NTU-MC page: `<http://compling.hss.ntu.edu.sg/ntumc/>`.

Missing senses are one of the major problems during the semantic annotation. We overcome this by integrating the wordnet editing interface provided by the OMW. Depending on the annotation task at hands, the annotation of a corpus can be done in parallel with the expansion of the respective wordnet's concept and sense inventory.

A third tagging interface (not shown) allows also the direct manipulation of the corpus structure. Its major features include creating, deleting and editing sentences, words and concepts. It is too generalized to be used as an efficient tagger, but it is useful to correct POS tags, tokenization errors and occasional spelling mistakes. It can also be used to correct or create complex concept structures of multi-word expressions, that could not be automatically identified.

The minimal input required by our interfaces (in the typical case, just clicking a radio button), especially the lexical tagger, ensures time isn't wasted with complex interfaces. It also guarantees that through the automated linking of the databases, we avoid typos and similar noise in the produced data. An earlier version allowed annotators to tag directly with synset IDs, but it turned out that it was very common for the ID to be mangled in some way, so we now only allow entering a synset through the linking to the OMW.

3.2 Annotation Agreement

IMI also includes a tool to measure inter-annotator agreement (Figure 3). Up to four annotations can be compared, for any section of the corpus. The tool also calculates the majority tag (MajTag). Average agreements scores are then computed between annotators and between annotators and the majority tag. Results are displayed by sentence and for the selected portion (e.g. the entire corpus). Agreement with the MajTag is color coded for each annotation so that the annotators can quickly spot disagreements. The interface provides quick access to database editing for all taggers, and to the OMW editing tools. The elected MajTag can also be automatically propagated as the final tag for every instance.

For some texts up to three annotators have been used, with one being a research assistant and two being students in a semantics class. These students only had a half hour of training, and used the sequential tagger to tag around 250 concepts each. The average inter-annotator agreement was 67.5%. Tagging speed was around 60 concepts/hour (self reported time). Note that roughly 25% of the potential concepts were pre-marked as `x`: entries such as preposition *in*, which should only be tagged on the very rare cases it is an adjective (*This is very in this year* or noun (*I live in Lafayette, IN*). Because the students were minimally trained (and not all highly motivated) we expected a low agreement. If two out of three annotators agreed then the words were tagged with the majority tag. Where all three annotators disagreed the students were required to discuss and re-tag those entries, and submit a report on them. An expert (the first author) then read (and marked) all the reports and fixed any tags where he disagreed with their proposed solution. Adjudicating and marking the reports takes about 30 minutes each, with some difficult to fix problems left for later. As a result of this process, all words have been seen by multiple annotators, and all hard ones by an expert (and our students have a much better understanding of the issues in representing word meaning using a fixed sense inventory)

For most texts, we only have enough funding to pay for a single annotator. **Targetted** tagging (annotating by word type) is known to be more accurate (Langone et al., 2004; Bond et al., 2008) and we use this for the single annotator. We expect to catch errors when we compare the annotations across languages: the annotation of the translation can serve as another annotator (although of course not all concepts match across languages).

3.3 Journaling

We take advantage of the relational database and use SQL triggers to keep track of every committed change, time-stamping and recording the annotator on every commit (true for both scripted and human manipulated data). The system requires going through a login system before granting access to the tools, hence permitting a detailed yet automatic journaling system. A detailed and tractable history of the annotation is available to control both the work-flow and check the quality of annotation. We can export the data into a variety of formats, such as RDF compatible XML and plain text triples.

3.4 Corpus Search Interface

Snapshots of the corpus are made available through an online corpus look up (Figure 4: available here: `<http://compling.hss.ntu.edu.sg/ntumc/cgi-bin/showcorpus.cgi>`). This search tool can query the corpus by concept key,

(SID:11140) **I gather from his wire that there have been some new incidents of importance."**

cid	lemma	A	B	C	D	MajTag	Comments
0	gather	00945125-v	00945125-v	00158804-v	00945125-v	00945125-v	
1	wire	06622709-n	06622709-n	None	06622709-n	06622709-n	
2	there	00109461-r	00109461-r	None	w	00109461-r	
3	have	x	x	x	x	x	**A:** auxiliary verb
4	be	02749904-v	x	None	02603699-v	?	
5	some	02267308-a	02267308-a	None	02267308-a	02267308-a	
6	new	00112601-r	02070491-a	None	00112601-r	00112601-r	
7	incident	07307477-n	07307477-n	01856929-a	07307477-n	07307477-n	
8	importance	05168261-n	05168261-n	05168261-n	05168261-n	05168261-n	
9	i	77000015-n	77000015-n	None	None	77000015-n	
10	his	77000050-n	77000050-n	None	None	77000050-n	
11	that	77000079-a	77000079-a	None	None	77000079-a	

	B	C	D	M
A	0.833	0.167	0.583	0.917
B	0.167	0.500	0.833	
C	0.167	0.167		
D	0.583			

Figure 3: Inter-annotator Agreement Reports

4 Results for: (C-Lemma:multi*)+(SID>=60000)

Sid	Sentence
60118	At the summit , the ministers will talk about diverse issues , including the development of an optical fiber network , standardization of information communications , and digitization of the public sector , aiming at the establishment of an information and communications infrastructure in the Asia-Pacific region in the **multi-media** age , and this summit is planned to be held annually afterwards , taking the initiative in the promotion of information technology in the region . 为了 完善 以 多媒体 时代 为 目标 的 亚太 信息 通信 基础 设施 ， 根据 对 扩充 光纤网 、 信息 通信 的 标准化 、 公共 领域 的 信息化 等 广泛 的 内容 进行 协商 的 方针 ， 今后 每 年 召开 例行 会议 ， 以 发挥 其 推进 [亚太 地区 信息化 指挥塔 的 作用 。 (cmn)
60246	" The WTO is the first organization which provides the basis for the stability and security of **multinational** trade . We must establish a trade system which works satisfactorily without squandering the trust in the GATT system . " “ WTO 是 第一 个 为 多边 贸易 提供 稳定 和 安全 基础 的 组织 。 它 必须 继续 保持 对 [关税 及 贸易 总协定 体制 的 信任感 ， 并 建立 能 充分 发挥 机能 的 贸易 体系 ” 。 (cmn)
60391	The union 's headquarters are in a one bedroom apartment in an old **multi-purpose** building in the Yaomade district of downtown Kowloon . 这个 协会 的 总部 设 在 [九龙 半岛 的 闹市 、 油麻 地区 的 一 栋 陈旧 的 杂居 公寓 里 的 一 居室 中 。 (cmn)
60766	The veil has not yet been lifted on the yacht and its **multitude** of modifications and improvements , but in terms of capabilities it certainly can not lose to any other boat . 虽然 这 艘 经过 多 次 改良 制成 的 帆船 尚 包裹 在 面纱 之中 ， 但 它 拥有 不 负于 任何 参赛 帆 船 的 能力 ， 这 一 点 是 确切无疑 的 。 (cmn)

Language: English | Chinese (simplified) Concept: C-lemma: multi* Word: Lemma:

SID (from): 60000 SID (to): Sentiment: MLSentiCon SentiWN POS: Select POS ? Limit: 10

Figure 4: Corpus Search Interface (results for the regular expression ‘`multi*`’ as concept lemma, using sentence ids to restrict the search to the Kyoto News Corpus, in bitext mode for Mandarin Chinese)

concept lemma, word, lemma, sentence id and POS, as well as any combination of these fields. Mousing over a word shows its lemma, pos, sense and annotators’ comments (if any), clicking on a word pops up more information about the lemma, pos and sense (such as definitions) that can be clicked for even more information. Further, it is possible to see aligned sentences (for as many languages as selected), and color coded sentiment scores using two freely available sentiment lexicons, the SentiWordNet (Baccianella et al., 2010) and the ML-SentiCon (Cruz et al., 2014) (individually or intersected). Further improvements will allow highlighting cross-lingual word and concept alignments (inspired by Nara: Song and Bond, 2009).

4 Summary and Future Work

We have described the main interfaces and functionality of IMI. It has undergone almost six years of development, and is now a mature annotation platform. The improvement of its interfaces and functionality have not only greatly boosted the speed of the NTU-MC annotation, but has also greatly facilitated its coordination - making it easier to maintain both consistency and quality of the corpus.

In the near future we intend to:

- refine the cross-lingual word and concept alignment tool (not shown here)
- develop a reporting interface, where the project coordinators can easily review the history of changes committed to the corpus database
- add a simple corpus import tool for adding new texts in different languages
- further develop the corpus search interface, to allow highlighting cross-lingual word and concept links
- implement more automated consistency checks (e.g. match lemmas of words with

the lemmas of concepts, verify that concept lemmas are still senses of the concept used to tag a word, etc.)

- improve graphical coherence, as different parts of the toolkit have originally been developed separately, as a whole, our system currently lacks graphical coherence

We hope that the open release of our system can motivate other projects to embrace semantic annotation projects, especially projects that are less oriented towards development of systems. We would like every wordnet to be accompanied by a sense-tagged corpus!

Acknowledgments

This research was supported in part by the MOE Tier 2 grant *That's what you meant: a Rich Representation for Manipulation of Meaning* (MOE ARC41/13). We would also like to thank our annotators for their hard work and patience during this system's development.

References

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. European Language Resources Association (ELRA), Valletta, Malta.

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 Ryan Foster. 2013. Linking and extending an open multilingual wordnet. In *51st Annual Meeting of the Association for Computational Linguistics: ACL-2013*, pages 1352–1362. Sofia. URL `http://aclweb.org/anthology/P13-1133`.

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, 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`.

Fermín L Cruz, José A Troyano, Beatriz Pontes, and F Javier Ortega. 2014. Building layered, multilingual sentiment lexicons at synset and lemma levels. *Expert Systems with Applications*, 41(13):5984–5994.

Luís Morgado da Costa and Francis Bond. 2015. OMWEdit - the integrated open multilingual wordnet editing system. In *ACL-2015 System Demonstrations*. (this volume).

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

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.

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

Svetla Koeva, Sv Leseva, and Maria Todorova. 2006. Bulgarian sense tagged corpus. In *Proceedings of the 5th SALTMIL Workshop on Minority Languages: Strategies for Developing Machine Translation for Minority Languages, Genoa, Italy*, pages 79–87.

Svetla Koeva, Borislav Rizov, and Svetlozara Leseva. 2008. Chooser: a multi-task annotation tool. In *LREC*.

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

Helen Langone, Benjamin R. Haskell, and George A. Miller. 2004. Annotating wordnet. In *Workshop On Frontiers In Corpus Annotation*, pages 63–69. ACL, Boston.

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.

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.

Rebecca J Passonneau, Ansaf Salleb-Aouissi, and Nancy Ide. 2009. Making sense of word sense variation. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, pages 2–9. Association for Computational Linguistics.

Sanghoun Song and Francis Bond. 2009. Online search interface for the Sejong Korean-Japanese bilingual corps and auto-interpolation of phrase alignment. In *Proceedings of the Third Linguistic Annotation Workshop (LAW III)*, pages 146–149. Singapore.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. brat: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations Session at EACL 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.

Shan Wang and Francis Bond. 2013. Building the Chinese Open Wordnet (COW): Starting from core synsets. In *Proceedings of the 11th Workshop on Asian Language Resources, a Workshop at IJCNLP-2013*, pages 10–18. Nagoya.