A Multilingual Sentiment Corpus for Chinese, English and Japanese

Francis Bond,* **Tomoko Ohkuma,**** **Luis Morgado Da Costa,***
Yasuhide Miura,** **Rachel Chen,*** **Takayuki Kuribayashi,*** **Wenjie Wang***
* **Nanyang Technological University**, Singapore
** **Fuji Xerox Corporation**, Japan
bond@ieee.org

Abstract

In this paper, we present the sentiment tagging of a multi-lingual corpus. The goal is to investigate how different languages encode sentiment, and compare the results with those given by existing resources. The results of annotating a corpus for both concept level and chunk level sentiment are analyzed.

Keywords: multilingual, sentiment, wordnet

1. Introduction

This paper present the results of annotating two English short stories (*The Adventure of the Speckled Band* and *The Adventure of the Dancing Men* (Conan Doyle, 1892, 1905)) and their Chinese and Japanese translations. We are currently expanding the annotation into more languages (starting with Indonesian) and more texts (software reviews). There are many corpora tagged for sentiment, for example the Stanford Sentiment Treebank (Socher et al., 2013), but few multilingual (Steinberger et al., 2011; Balahur and Turchi, 2014) and no multilingual sentiment corpora for Asian languages. (Prettenhofer and Stein, 2010) contains English, French, German and Japanese product reviews, but they are comparable (reviews of the same product) or machine translated, not translated text, so while useful it is not suitable for studying close correspondences.

2. The Corpus

To compare the expression of sentiment in Chinese, English and Japanese, we used text from the NTU Multilingual Corpus (Tan and Bond, 2012). The corpus was already tagged with concepts (synsets) using the open multilingual wordnet (Bond and Foster, 2013). The entries for the three languages are based on the Princeton Wordnet for English (Fellbaum, 1998), the Chinese Open Wordnet for Chinese (Wang and Bond, 2013) and the Japanese wordnet for Japanese (Bond et al., 2009). In addition, we added pronouns (Seah and Bond, 2014) and new concepts that appeared in the corpus. We also have translations for *The Adventure of the Speckled Band* in Bulgarian, Dutch, German, Indonesian and Italian, and are in the process of expanding the annotation.

We chose a literary text, because we are interested in how sentiment is used in building a coherent narrative. We wish to consider questions such as how different characters are portrayed, whether sentiment follows the structure of the story and if translators prefer words with the same literal meaning or the same connotation.

Annotation was done using **IMI** — A Multilingual Semantic Annotation Environment (Bond et al., 2015), extended to allow for the annotation of sentiment at concept and chunk level. We use a continuous scale for tagging sentiment, with scores from -100 to 100. The tagging tool splits these into seven values by default (-95, -64, -34, 0, 34, 64, 95), and there are keyboard shortcuts to select these values. Annotators can select different, more fine-grained values if they desire. The annotators were told to tag using several evaluative adjectives as guidelines, shown in Table 1. The table also shows new examples from the corpus after annotation.

Each of the three texts was annotated by a single native speaker for that language, then the different languages were compared, major differences discussed and, where appropriate, retagged. If they were not sure whether the text segment shows sentiment or not, annotators were instructed to leave it neutral (0).

3. Concept Level Annotation

At the lexical level, we annotate concepts (words that appear in wordnet) that, in context, clearly show positive or negative sentiment. Operators such as ***very*** and ***not*** were not tagged. Concepts can be multiword expressions, for example *give rise* "produce" or *kuchi-wo hiraku* "speak". Each corpus was annotated with a single annotator with linguistic training.

The size of the corpus is shown in Table 2. English is the source language, the translators have separated some long sentences into shorter ones for both Chinese and Japanese. Chinese words are in general decomposed more than English, and the wordnet has fewer multi-word expressions so the corpus has more concepts. Japanese has no equivalent to some common concepts such as ***be*** in *I am happy*, and drops the subject when it is clear from the context and thus has many fewer concepts.

Ideally, multiple annotators for each language would give even more reliable results, but we decided to use a single annotator for the following reasons. The first is that the corpus has already been annotated for sense

Score	Example	Example	Example	Corpus Examples
95	fantastic	very good		perfect, splendidly
64	good	good		soothing, pleasure
34	ok	sort of good	not bad	easy, interesting
0	beige	neutral		puff
-34	poorly	a bit bad		rumour, cripple
-64	bad	bad	not good	hideous, death
-95	awful	very bad		deadly, horror-stricken

Table 1: Guidelines for sentiment score given to annotators

Language	Sentences	Words	Concepts	Distinct Concepts
English	1,199	23,086	12,972	3,494
Chinese	1,225	24,238	16,285	3,746
Japanese	1,400	27,408	10,095	2,926

Table 2: Size of the Corpus for the three languages

(Bond et al., 2013) and therefore the annotators have more information about the individual lexical items available to them. Secondly, we compare the annotation across the languages: if we consider the translations as one corpus, then we are annotating three times and we do compare the annotator agreement (§ 3.1.). Finally, there is the question of cost: we only had enough money to pay three annotators, and wanted to have data in three languages.

The first of our quality control measures was to look at words both in context and then out of context. After the initial annotation (done sentence-by-sentence), the annotators were shown the scores organized per word and per sense: where there was a large divergence (greater than one standard deviation), they went back and checked their scores.

Some examples of high and low scoring concepts and their lemmas are given in Table 3. The score for the concept is the average over all the lemmas in all the languages. The concepts are identified with the Interlingual Index Bond et al. (2016).[1]

3.1. Cross-lingual Comparison

In this section we take a look at the agreement across the three languages. We examined each pair (Chinese-English, Chinese-Japanese and English-Japanese), and measured their correlation using the Pearson product-moment correlation coefficient (ρ), as shown in Table 4. This was calculated over all concepts which appeared in both languages. Because translations are not one-to-one, we matched concepts, and took the average sentiment score per language, repeated as often as the minimum frequency in both languages. Thus for example, if between Chinese and English, `02433000-a` "showing the wearing effects of overwork or care or suffering" appeared three times in Chinese (as 憔悴 *qiáo cuì*) with an average score of -48.5 and twice in English with a score of -64 (as *haggard* and *drawn*), we would count this as *two* occurrences of -48.5 (in Chinese) and -64 (in English). In general, fewer than half of the concepts align across any two languages (Bond et al., 2013).

Pair	ρ	# samples
Chinese-English	.73	6,843
Chinese-Japanese	.77	4,099
English-Japanese	.76	4,163

Table 4: Correlation between the different language pairs

For most concepts, the agreement across languages was high, although rarely identical. There was high agreement for the polarity but not necessarily in intensity/magnitude. For example, for the concept `02433000-a` "haggard", the English words *drawn* and *haggard* were given scores of -64, while Chinese 憔悴 was given only -34.

An example of different polarity was the English lemma "great" for synset `01386883-a`, which received a score of 45.2, whereas the Japanese lemma 大きい for the same synset received a score of 0 (neutral).

In addition, lemmas in the same synset might have another sense that is positive or negative, and this difference causes them to be perceived more or less positively. For example, in English, both *imagine* and *guess* are lemmas under synset `00631737-v`, but *imagine* is perceived to be more positive than *guess* because of their other senses. This cross-concept sensitivity can differ from language to language, thus causing further differences. In general, the English annotator was more sensitive to this, which explained much of the difference in the scores. Overall, cross-lingual comparisons of concepts that were lower in agreement were due to both language and annotator differences. The English annotator had generally been more extreme in the rating compared to the Chinese and Japanese annotators.

[1] LOD: `http://www.globalwordnet.org/ili/ixxx`.

Concept	freq	score	English	score	Chinese	score	Japanese	Score
i40833	24	+50	marriage	39	婚事	34	結婚	58
			wedding	34				
i11080	5	+40	rich	33	有钱	34	裕福	66
i72643	4	+33	smile	32	微笑	34	笑み	
i23529	40	−68	die	−80	去世	−60	亡くなる	−63
					死亡	−64	死ぬ	−62
i36562	5	−83	murder	−95	谋杀	−95	殺し	−64
							殺害	−63

Table 3: Examples of high and low scoring concepts, only total frequencies shown.

3.2. Comparison with Sentiwordnet and MLSentiCon

We also measured agreement with the widely used Sentiwordnet (Baccianella et al., 2010) and the newer MLSentiCon (Cruz et al., 2014), both of which are automatically-generated resources. Here, we compared at the synset level, comparing all concepts that appeared at least once in any language, averaged over all occurrences in all three languages. So for the example given above, the score would be 54.7. The results are given in Table 5. Here we are measuring over distinct concepts, with no weighting. For the sentiment lexicons, we give results over the subset in the corpus, and over all synsets.

Pair	ρ	**# samples**
SentiWN-MLSenticon	.51	6,186
	.42	123,845
NTUMC-SentiWN	.42	6,186
NTUMC-MLSenticon	.48	6,186

Table 5: Correlation between the different resources

The results show that none of these three resources agree very well. The automatically created resources related better with each other, but still had a low correlation. Neither resource closely correlated with the examples seen in context in the corpus: the newer MLSenticon having slightly better agreement.

Examining the examples by hand, many concepts we marked as neutral received a score in these resources (e.g. *be* which is +0.125 in Sentiwordnet or *April*, which is -0.125 in MLSenticon), while other concepts for which we gave a strong score (e.g *violence* -64) were neutral in these other resources. As our senses were confirmed by use in a corpus, we consider our scores to be more accurate.

Sentiwordnet and MLSentiCon were both produced by graph propagation from a small number of seeds (around 14). It would be interesting to try to add our new data (suitably normalized) as new seeds and try to recalculate the scores: a larger pool of seeds should give better results.

4. Chunk Level Annotation

In this phase we tagged larger units. The goal is to tag groups of words, that at a given level share the same polarity and intensity. Here we include the effects of operators. In order to reduce effort, we do not mark all chunks, but only those where the polarity or strength changed. We always give the sentence (the largest possible chunk) a score.

We give some (artificial) examples below (taken from the tagging guidelines).

(1) I think this is very good

+64	good
+95	very good
+95	this is very good
+90	I think this is very good

(2) Do you think this is very good?

+64	good
+95	very good
+95	this is very good
+0	Do you think this is very good?

(3) The horse raced past the barn.

+0	The horse raced past the barn.

(4) I do not understand.

+33	understand	
-33	not understand	polarity change
-33	I do not understand	

We compared the sentence level annotation across languages in Table 6, and found the agreement less good than for concepts, but still generally ok. The majority of sentences were neutral. The annotators found this task hard to do, especially deciding on chunk boundaries.

Pair	ρ	**# samples**
English-Chinese	.60	1,084
English-Japanese	.56	873
Chinese-Japanese	.70	713

Only for sentences that aligned one-to-one.

Table 6: Cross-lingual Sentence Correlation

Corpus examples

We look at two Mandarin Chinese examples from the actual tagged corpus, demonstrating how sentiment changes value with the effects of operators. As we see in (6), a negative operator does not necessarily just flip the sentiment score, it may also effect the value.

(5) 没有 表示 异议
méi-yǒu biǎo-shì yì-yì
not-have indicate objection

"did not object"		
-34	异议	
-34	表示 异议	
+34	没有 表示 异议	polarity change

(6) 决 不 反对
jué bù fǎn-duì
certainly not object

"certainly not object"		
-34	反对	
+15	不 反对	polarity change
+34	决 不 反对	intensity change

An area which is currently not indicated in the sentiment rating are devices which operate at a layer above the surface chunk, such as sarcasm. Sarcasm, in most cases, could cause another flip in polarity. At present, we chose to indicate such instances in the comments (e.g. "SARCASM"), but otherwise leave the sentiment rating as-is. In fact, the stories we annotated did not have any examples of sarcasm or irony.

5. Discussion and Future Work

In this paper we presented an initial multilingual annotation for sentiment at the lexical and chunk level over Chinese, English and Japanese languages. These results show that sentiment, at the lexical level, can be modelled with concepts that retain their scores across languages. We can thus produce a good first annotation by sense-tagging and then adding sentiment.
In future work, we want to model and annotate (i) the effects of operators and (ii) the targets of the sentiment, as well as expand the corpus to cover more text in more languages.

Acknowledgments

This research was partially supported by Fuji Xerox Corporation through joint research on *Multilingual Semantic Analysis*.

References

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. European Language Resources Association (ELRA).

Alexandra Balahur and Marco Turchi. 2014. Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis. *Computer Speech & Language*, 28(1):56–75.

Francis Bond, Luís Morgado da Costa, and Tuấn Anh Lê. 2015. IMI — a multilingual semantic annotation environment. In *ACL-2015 System Demonstrations*.

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, Hitoshi Isahara, Sanae Fujita, Kiyotaka Uchimoto, Takayuki Kuribayashi, and Kyoko Kanzaki. 2009. Enhancing the Japanese WordNet. In *The 7th Workshop on Asian Language Resources*, pages 1–8. ACL-IJCNLP 2009, Singapore.

Francis Bond, Piek Vossen, John McCrae, and Christiane Fellbaum. 2016. CILI: The collaborative interlingual index. In *Proceedings of the 8th Global Wordnet Conference (GWC 2016)*. (submitted).

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`.

Arthur Conan Doyle. 1892. *The Adventures of Sherlock Homes*. George Newnes, London.

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

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*.

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

Peter Prettenhofer and Benno Stein. 2010. Cross-Language Text Classification using Structural Correspondence Learning. In *48th Annual Meeting of the Association of Computational Linguistics (ACL 10)*, pages 1118–1127. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/P10-1114`.

Yu Jie Seah and Francis Bond. 2014. Annotation of pronouns in a multilingual corpus of Mandarin Chinese, English and Japanese. In *10th Joint ACL - ISO Workshop on Interoperable Semantic Annotation*. Reykjavik.

Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *EMNLP*.

Josef Steinberger, Polina Lenkova, Mijail Alexandrov Kabadjov, Ralf Steinberger, and Erik Van der Goot. 2011. Multilingual entity-centered sentiment analysis evaluated by parallel corpora. In *RANLP*, pages 770–775. Citeseer.

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.