A Comparison of Sense-level Sentiment Scores

Francis Bond,[♠] Arkadiusz Janz[♢] and Maciej Piasecki[♢]
[♠] Nanyang Technological University, Singapore
[♢] Wrocław University of Science and Technology, Poland
bond@ieee.org, {arkadius.janz|maciej.piasecki}@pwr.edu.pl

Abstract

In this paper, we compare a variety of sense-tagged sentiment resources, including SentiWordNet, ML-Senticon, plWordNet emo and the NTU Multilingual Corpus. The goal is to investigate the quality of the resources and see how well the sentiment polarity annotation maps across languages.

1 Introduction

There are several semantic resources with senses annotated by sentiment polarity, e.g. SentiWordNet 3.0 (Baccianella et al., 2010) and even emotions, e.g. WordNet-Affect (Strapparava and Valitutti, 2004; Torii et al., 2011). However, most of them were built on the basis of automated expansion of a small subset of senses described manually. In addition the majority of them were built for a single language, namely English, with ML-SentiCon (Cruz et al., 2014) a notable exception.

This paper present the results of comparing two very different sense-level sentiment resources: a very large semantic lexicon annotated manually for Polish, i.e. plWordNet (Maziarz et al., 2016) expanded with manual emotive annotations (Zaśko-Zielińska et al., 2015); the annotation of 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 (Bond et al., 2016a). As the stories have been annotated on the basis of senses, not words – i.e. all words were assigned Princeton WordNet synsets – this opens an unique possibility of cross-lingual comparison of manual sentiment annotation at the level of word senses. These are then compared with SentiWordNet and ML-SentiCon and finally they are all compared to a small gold standard sample Micro-WNOp Corpus (Cerini et al., 2007).

Our technical goal is to analyse the feasibility and technical means of correlation between independently created resources as the first step towards cross-lingual applications. Taking a more fundamental perspective, we want to investigate the level and distribution of correlation between sentiment polarity expression on the sense level between languages. In addition this is also an exercise in utilisation of the interlingual manual mapping between plWordNet and Princeton WordNet that has been built independently.

2 Resources

In this section we describe the resources we used.

2.1 SentiWordNet

SentiWordNet (Esuli and Sebastiani, 2006) annotates a synset with three numerical values in the range $\langle 0, 1 \rangle$ placing the synset in a three dimensional polarity space. The dimensions describe "how objective, positive, and negative the terms contained in the synset are". As the three values must sum to one, there are only two degrees of freedom.

About 10% of the adjectives were manually annotated, each by 3-5 annotators (Baccianella et al., 2010). In SentiWordNet 3.0, the automated annotation process starts with all the synsets which include 7 "paradigmatically positive" and 7 "paradigmatically negative" lemmas.[1] The initial seed is expanded with a random walk algorithm to generate a training set for a committee of classifiers and estimate final polarity scores of synsets. In the end, SentiWordNet 3.0 added automatic sentiment annotation to all of Princeton WordNet 3.0.

[1] good, nice, excellent, positive, fortunate, correct, superior; bad, nasty, poor, negative, unfortunate, wrong, inferior (Turney and Littman, 2003)

2.2 ML-SentiCon

The method proposed in Baccianella et al. (2010) has become the motivation for further work on the development of word-level and sense-level sentiment lexicons. ML-SentiCon (Cruz et al., 2014) expands the idea presented in (Baccianella et al., 2010) by introducing additional sources of information such as WordNet-Affect (Strapparava and Valitutti, 2004) and General Inquirer (Stone et al., 1966) to improve the accuracy and coverage of initial polarity seed. The seed is expanded using the same general approach proposed in Baccianella et al. (2010). However, instead of a single score for each synset, individual scores for each sense are calculated, and then synset scores are calucalted by averaging these.

2.3 plWordNet 4.0 emo

In plWordNet the emotive annotation is assigned not to synsets, but to senses (also known as lexical units: LU), i.e. pairs of lemmas and synsets. These are represented internally as triples of lemma, Part of Speech and sense identifier (number) – every sense belongs to exactly one synset, so a synset represents a sense – a lexical meaning. Senses are fundamental elements of the plWordNet structure, cf (Maziarz et al., 2016).

From the point of view of emotional sentiment polarity, plWordNet senses are divided into *marked* and *neutral*. The first can be also called *polarised*. Polarised senses are assigned the *intensity* of the sentiment polarisation, *basic emotions* and *fundamental human values*. The latter two provide additional characteristics and help annotators to determine the sentiment polarity and its intensity expressed in the 5 grade scale: *strong* or *weak* vs *negative* and *positive*. Each annotator's decision for polarised senses is supported by use examples – a sentence including the given sense and illustrating the postulated sentiment polarity and its strength.

Concerning emotions, due to the compatibility with other wordnet-based annotations, the set of eight basic emotions recognised by Plutchik (Plutchik, 1980) were used (Zaśko-Zielińska et al., 2015). It contains Ekman's six basic emotions (Ekman, 1992): *joy* , *fear*, *surprise*, *sadness*, *disgust*, *anger*, complemented by Plutchik's *trust* and *anticipation*. As a result, negative emotions do not prevail in the set. One sense can be assigned more than one emotion and, as a result, complex emotions can be represented by using the same eight-element set, following the observations of Plutchik (1980).

However, as the comparison we aim for is limited only to sentiment polarity, both emotions and fundamental values will be ignored in comparison.

2.4 NTU Multilingual Corpus

The NTU Multilingual Corpus (Tan and Bond, 2012) has a variety of texts and their translations, many of which are sense annotated.[2]

Two stories from the Sherlock Holmes Canon (*The Adventure of the Speckled Band* and *The Adventure of the Dancing Men*) have been both sense tagged with wordnet senses and annotated for sentiment (Bond et al., 2016a). Princeton Wordnet (Fellbaum, 1998) was used for English, the Chinese Open Wordnet for Chinese (Wang and Bond, 2013) and the Japanese wordnet for Japanese (Bond et al., 2009). These are linked through Princeton WordNet 3.0 (Fellbaum, 1998) with the help of the open multilingual wordnet (Bond and Foster, 2013). In addition, pronouns (Seah and Bond, 2014) and new concepts that were discovered in the corpus during the annotation have been added.

A continuous scale was used 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. Three values were chosen for each polarity, in order to be able to show the changes in chunks: *quite good* is less positive than *good* and this is less positive than *very good*. Annotators could select different, more fine-grained values if they desire. The annotators were given several exemplars as guidelines, shown in Table 1. The final column of the table shows 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 untagged.

In this paper, we only use the sense level annotation, and ignore chunks. Like plWordNet emo, only marked senses are annotated: those senses of

[2]The corpora are searchable here: `http://compling.hss.ntu.edu.sg/ntumc/`. They will be made avaialble for download by the time of the conference.

Score	Example	Chunk 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: Exemplars for sentiment scores

words in text that, in context, clearly show positive or negative sentiment were annotated. If a sense is not annotated, then we treat it as an implicit tag of neutral (zero). Operators such as ***very*** and ***not*** were not tagged. Concepts can be multiword expressions, for example *give rise* "produce" or 口を開く *kuchi-wo hiraku* "speak". Each corpus was annotated by a single annotator with linguistic training.

Lang.	Sent.	Words	Concepts	Distinct
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

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.

There was some quality control: senses were examined 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., 2016b).[3]

2.5 The Micro-WNOp Corpus

We evaluated the Micro-WNOp Corpus (Cerini et al., 2007) as it is the only **sense-tagged** sentiment lexicon we could find.[4] It was used to evaluate SentiWordNet and build ML-SentiCon, and consists of 1,105 Wordnet synsets chosen from the General Inquirer lexicon (Stone et al., 1966) and annotated by 1–3 annotators.

There are many corpora tagged for sentiment, for example the Stanford Sentiment Treebank (Socher et al., 2013), but few multilingual (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.

3 Comparisons

We are going to compare four languages and two types of resources: a corpus and a lexicon from the perspective of sentiment polarity annotation. In order to make the comparison feasible, we focus on word senses – that can be represented by concepts – and their mappings across languages, as links between the different resources. There are both manually annotated and automatically built (to a very large extent) resources among the compared ones. Finally two types of the sentiment polarity annotations that are represented by the compared resources use similar but slightly different models: the semi-continuous scale, e.g. NTU-MC and the discrete scale, e.g. the five-grades scale of plWordNet emo.

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

[4] `http://www-3.unipv.it/wnop/`

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 from NTU-MC, only total frequencies shown.

3.1 Cross-lingual Comparison inside the Corpus

In this section we take a look at the agreement across the three languages of the NTU-MC. 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. We chose this as it is invariant under separate changes in location and scale. This was calculated over all concepts which appeared in both languages. All three wordnets (Sec. 2.4) use the same conceptual structure, that of Princeton Wordnet. When we compare, it makes no sense to compare senses, as they are language specific.

Instead, we matched concepts, represented by synsets. For each language, we calculated the sentiment score for a synset by averaging over all its senses. When we compare across languages, if a synset appears in the corpus multiple times, we add it to the comparison set as often as the least frequent language. 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 directly across any two languages (Bond et al., 2013). Even though we have over 12,000 occurences concepts in English and more in Chinese and Japanese (Table 2) fewer than 7,000 appear in both (Table 4).

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 *qiáo cuì* was given a weaker score of -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.

3.2 Cross-lingual Comparison: Corpus vs Wordnet

NTU-MC and plWN have different sentiment annotation schema. The first one allows for a scale close to continuous: $\langle -100, +100 \rangle$, while the latter uses only 5-degree polarity scale (including *neutral*). In practice, most senses are annotated using the default values, which groups the scores around seven points: three positive and three negative.

NTU-MC annotation was done on the level of word senses represented by PWN synsets. The mapping between plWN and PWN is defined on the level of synsets. Thus, first both annotations in both resources, namely, NTU-MC and plWN had to mapped onto the level of synsets. In the case of NTU-MC we applied the same strategy as above: every synset is assigned a polarity score which is the average across the polarity values assigned to its senses in the corpus (respectively to a given language under examination). This procedure introduces an implicit weighting: more frequent senses have bigger influence on the synset polarity. In addition the polarity values do not need to be constant for a given sense in all its occurrences. So, by averaging them for one synset we additionally balance between small differences resulting from different contexts.

The scale in plWordNet is discrete and semi-continuous in NTU-MC.[5] As any attempt to make the plWordNet scale continuous would be arbitrary (only one dimension and up to three annotations per a sense), we decided to map the NTU-MC scale onto a discrete set of values, namely the five degree scale of plWordNet. First, we generated a histogram of averaged polarity values in which we could observe quasi-Gaussian concentrations of values around ± 34. On the basis of the distribution of values in the histogram we defined thresholds for weak polarity on ± 17. In the case of higher (or lower) polarity of synsets in NTU-MC we could notice that two maxima located around ± 64 and ± 95 were not significantly separated between them, while very distinctively separated from the first one. Thus we decided to treat them as representing one category of strong positive/negative polarity and to set up the threshold for them on ± 54.

In plWordNet in order to obtain synset polarity scores on the basis of sense scores, we cannot simply average them, as the scale consist of only two levels (in each direction) and the average number of senses in a synset is below 2. Thus, the synset polarity is obtained on the basis of simple majority voting[6] from the sense values. In case of a tie, we take the maximum or minimum value, respectively for positive and negative.

In order to identify the corresponding plWordNet and Princeton WordNet synsets, we utilised the manually constructed mapping between both wordnets. It is based on different inter-lingual relations that link synsets and express different levels and forms of meaning correspondence from the very strong correspondence in the case of *I-synonymy* (interlingual) down till, e.g., *I-holonymy* which signals that the target represents a whole that includes the part represented by the source. The mapping procedure organises the inter-lingual relations into a kind of decision lists (one for each Part of Speech) that guide linguists from the strongest relations – also the most informative – to the weakest. The idea was to not leave any synset not mapped, even if only some weak form of correspondence can be expressed. Due to the different types of inter-lingual meaning correspondence, we expected also different levels of correlation between sentiment annotations assigned to the mapped synsets. On the basis of the properties of the inter-lingual relations and the mapping decision lists we divided I-relations into four groups: *synonymic*, *hyponymy*, *hypernymy* and *other*. The first group encompasses *I-synonymy*, *I-partial-synonymy* and *I-interparadigmatic-synonymy* (restricted to *adj–adv* links only).

I-hyponymy is most numerous relation, and expresses that the source synset has more narrow meaning, but mostly it is very close to the meaning represented by the target. The group was extended with *I-inter-register-synonymy* links which share similar properties to *I-hyponymy* links in terms of meaning and polarity.

I-hypernymy is used when the synset of the source wordnet (for which the mapping is built) represents more general meaning than the synset of the target wordnet, so it is a reverse relation to *I-hyponymy*. However, I-hypernymy is further in the mapping decision list than I-hyponymy, so it is used in less clear mapping situations and expresses

[5] I.e. *de facto* discrete on the level of senses and more continuous after averaging

[6] plWN annotation include about 5% of ambiguous senses that can express in some contexts positive or negative polarity. For them both values are taken into account during voting.

significantly weaker correspondence.

The *other* category groups all the rest of inter-lingual relation that are used mostly as a last resort mapping decision, so they signal weak meaning correspondence.

For the comparison we used two different measures. Firstly, in a similar way like for inter-lingual comparison in NTU-MC (see the previous section), we calculated *Pearson's correlation* of the synset scores, setting plWordNet emo's weak to 0.4 and strong to 0.8.

Secondly, we discretised NTU-MC synset (concept) scores to the five grade scale of plWordNet (following the procedure described earlier) and checked the agreement between the resulting values with the sentiment polarity values of plWordNet synsets. Cohen's κ was used to measure the agreement.

The results of both types of comparison are presented in Table 5. The ρ column presents the measured correlation. As sentiment annotations are quite remotely related to each other (done on the level of senses, for two languages, mapped by inter-lingual relations etc.), we decided to measure the agreement in two versions: κ_1 – only the sign of polarity (negative, neutral and positive), and κ_2 – five grade scale. The last column – **#synsets** – tells for how many Polish synsets we managed to establish links to the the synsets annotated in NTU-MC.

Type	ρ	κ_1	κ_2	**#synsets**
Synonymic	**.65**	.60	.53	1,043
Hyponymy	.62	.53	.47	1,271
Hypernymy	.56	.50	.33	147
All links	.63	.55	.48	1,880

Table 5: Correlation and Cohen's Kappa for matched annotations with respect to a type of inter-lingual connection between plWordNet and NTU-MC.

The correlation and agreement are the highest for the synonymic group of inter-lingual relations, as we could expect. The correlation does not drop much for the I-hyponymy group, but the agreement for both non-synonymic relations is significantly lower.

We do not provide results for the *other* category of mapping relations, as we could detect only a small number of links.

Concerning the agreement, it appeared to be good when only the polarity sign is concerned (κ_1), and it is still positive in the case of the full five grade scale (κ_2). The use of the hyponymy and hypernymy categories of links resulted also in a significantly lower, but still positive agreement. All three measures showed continuously decreasing and lower agreement when we apply less and less informative inter-lingual relations.

3.3 Cross-lingual Comparison: Analysis of Discrepancies

Limited agreement between the two manual resources means that there must large number of differences in annotations. In order to understand better the nature of these discrepancies we took a closer look into them into comparisons based on the synonymic inter-lingual relations. Most of the differences in this category result from different levels of the polarity. Only 5.6% of them express significant disagreement, i.e. different sign of polarity. One other co-authors has manually surveyed them to find that there are only 14 cases of two opposite polarity values, and a larger number of cases in which neutral polarity (i.e. the lack of polarity) on one side is mapped on the marked polarity on the other side (67,6%). Concerning the first, the strongest difference type, all such cases are listed in Table 6.

Sense	Pl	MC	Cause
incredible.1 (adj)	-s	+w	err. in plWN
extreme.1 (adj)	-s	+w	more narrow Polish meaning
impassable.1 (adj)	-w	+s	err. in NTU-MC
crazy.2 (adj)	+w	-s	I-part-syn.
grave.1 (adj)	+s	-s	err. in plWN
flare-up.1 (verb)	+s	-w	too strong I-relation
attack.5 (verb)	+w	-s	err. in plWN
blackguard.1 (verb)	-w	+s	err. in NTU-MC
fancy.1 (noun)	-w	+s	err. in NTU-MC
glimmer.2 (noun)	+w	-w	err. in both

Table 6: Survey of the strongest differences between the annotation of plWordNet and NTU-MC, where s = strong, w = weak.

As we could notice in Table 6, there is very little disagreement for nouns, only for adjectives and verbs that are much more difficult for both inter-lingual mapping and emotive annotation. The vast majority of disagreements resulted from the errors in the original annotations, e.g.: *incredible.1* – on

the Polish side the emotive annotation is based on wrong sense interpretation; *extreme.1* – the corresponding Polish sense was interpreted in a more narrow way, with a tendency to negative interpretation of extreme; *impassable.1* – a very likely error in NTU-MC error, it is hard to imagine a positive interpretation of this sense on the basis of the examples from the corpus, etc. The other two discrepancies seem to be caused by the mapping with the help of I-partial-synonymy. It expresses overlapping meaning, so their overlaps do not need to match the assigned sentiment annotations.

For *glimmer.1* it appears as *gleam* in *"See here, mister!" he cried, with a gleam of suspicion in his eyes, "you're not trying to scare me over this, are you?*. The complement *suspicion* is clearly negative but *gleam* is probably neutral, neither resource was perfect, and may have been biased by the context.

We also examined disagreements that involve neutral annotations: that is, in one resource the score is neutral (zero) and in the other is carries sentiment. In almost all cases, the neutral score was wrong. Annotators in NTU-MC were allowed to omit explicit neutral annotation and leave words unannotated in such cases. This resulted in some number of mistakenly skipped words. In a similar way, the vast majority of plWordNet:neutral vs NTU-MC:polarised cases is the combined result of gaps in the plWordNet sentiment annotation and a default rule that all gaps should be treated as neutral cases. The annotation was done for almost 90,000 senses, but this is around half of the wordnet. The default rule works quite well for nouns, where potentially neutral hypernymy branches were intentionally excluded from annotation, but fails definitely for other Parts of Speech.

3.4 Comparison with SentiWordNet and ML-SentiCon

Next, we compared both manually annotated resources, namely, plWordNet and NTU-MC with two resources used in many applications: SentiWordNet (Baccianella et al., 2010) and the newer ML-SentiCon (Cruz et al., 2014), discussed shortly in Sec. 2.1 and 2.2. As it was already mentioned, the sentiment annotation in both these resources were automatically propagated from a small set of manually prepared seeds.

SentiWordNet and ML-SentiCon are annotated on the level of synsets, so we used exactly the same pre-processing of plWordNet and NTU-MC. In the case of plWordNet we used also the same inter-lingual relation to map the Polish synsets onto Princeton WordNet ones. The Pearson's correlation for polarity values is presented in Table 7. 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
plWN – SentiWN	.32	22,435
plWN – MLSenticon	.41	22,435
plWN – NTUMC	**.63**	1,880

Table 7: Correlation between the different resources

The results show that none of these four resources agree very well. The automatically created resources related better with each other, but still had a low correlation. Their correlation is significantly smaller than the manually annotated NTU-MC and plWordNet. That is even more significant, when we take into account that the manually annotated resources were created for different languages, are based on different annotation models and we required the help of inter-lingual relations to map them. This whole process had to hamper the observed correlation. Neither automatically built resource closely correlated with the examples seen in context in the corpus and in the plWordNet use examples. However, the newer ML-SentiCon has 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 ML-SentiCon), 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 manual inspection, we consider our scores to be more accurate.

SentiWordNet and ML-SentiCon were both produced by graph propagation. SentiWordNet from a small number of seeds (around 14) and ML-SentiCon from more. It would be interesting to try to add our new data (suitably normalised) as new

seeds and try to recalculate the scores: a larger pool of seeds should give better results.

3.5 Evaluation with the Micro-WNOp Corpus

The Micro-WNOp Corpus was chosen to evaluate our resources, as it is commonly used and well balanced. First, we calculated the agreement for different annotators in the corpus. In group 1, with three annotators, we calculated annotator one vs the average of two and three, then two vs one and three and three vs one and two (ρ = 0.85, 0.78, 0.83 respectively, mean is 0.82). For group 2 with two annotators we compared them to each other (ρ = 0.94). In each case, we summed positive and negative to get a single score and compared using the Pearson product-moment correlation (ρ). This give us an upper bound for human agreement.

Both plWordNet and NTU-MC have far higher correlations than SentiWN, although with no results for many synsets. This shows the well known effect that hand-built resources are more reliable, but generally sparser.

Pair	ρ	# syn.
Micro-WNOp InterAnnotator	.88	995
Micro-WNOp – plWN	.77	413
Micro-WNOp – NTU-MC	.75	130
Micro-WNOp – SentiWN	.63	1,048
Micro-WNOp – plWN&NTU-MC	**.78**	352

Table 8: Correlation of Micro-WNOp lexicon with other resources

For completeness, we also calculated the correlation between Micro-WNOp and ML-SentiCon $\rho = .96$. However, as Micro-WNOp was used to as training data for ML-SentiCon the evaluation is not meaningful and we do not include it in Table 8.

4 The combined sentiment lexicon

One clear results of this comparison is that comparing the lexicons with each other improves them. Places where there was a difference in polarity or in zero vs non-zero sentiment were almost all errors. Once discovered there are easy to fix, and we have shared the results with the resource creators. Because the scores are different (a continuous score for NTU-MC and a 5 point scale for plWordNet emo) we can combine in two ways: binning NTU-MC or setting values for weak and strong for plWordNet emo (we used 0.4 and 0.8). They can then be combined over all synsets, to give a single resource that should be somewhat more accurate then either alone.

To combine the lexicons we decided to use binning strategy on NTU-MC and Micro-WNOp followed by a simple selection procedure. To represent matched concepts within the same category set we used thresholding function with thresholds being a result of score distribution analysis. In case of NTU-MC the following bins were proposed: $|s| \leq 0.18$ for neutral category, $0.18 < |s| \leq 0.54$ for weak polarity and $|s| > 0.54$ for strong polarity. First we selected a subset of paired synsets annotated both in NTU-MC and plWordNet emo which were compatible in terms of their polarity categories. To reduce the discrepancy between the annotations we also decided to remove all of paired synsets having different polarity categories. In the last step we introduce a group of unmatched synsets with their annotations to extend the coverage of joint lexicon. The final lexicon was evaluated again on Micro-WNOp (Table 7) giving a slight improvement of correlation.

5 Conclusion and Future Work

In this paper we presented a comparison of wordnet-based sense-level sentiment lexicons. We showed that the two manually annotated resources were more accurate than the semi-automatically created resources. We also showed that linking across languages preserved most of the valence (ρ = 0.65 – 0.77 for equivalent synsets). This means that the resources can be used for other languages, linked either directly or through an interlingual index. Finally we showed how they could be improved further by cross-checking and resolving inconsistencies, or by combining them.

In future work, we will: (i) correct the errors in the two resources and recalculate their correlation (as it is sensitive to outliers). (ii) create further sense-annotated sentiment tagged text

- Another Sherlock Holmes story (*The Red-Headed League*)
- Other translations for *The Adventure of the Speckled Band*: we have Bulgarian, Dutch, German, Indonesian, Italian and Polish, and are in the process of annotating them.

and (iii) model the effects of operators on lexemes to allow for compositional changes.

Acknowledgments

This research was partially supported by Fuji Xerox Corporation through joint research on *Multilingual Semantic Analysis*, CLARIN, the EU RISE Project 691152 — RENOIR *Reverse EngiNeering of sOcial Information pRocessing* and the NTU Digital Humanities Research Cluster.

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.

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 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, Tomoko Ohkuma, Luís Morgado da Costa, Yasuhide Miura, Rachel Chen, Takayuki Kuribayashi, and Wenjie Wang. 2016a. A multilingual sentiment corpus for Chinese, English and Japanese. In *6th Emotion and Sentiment Analysis Workshop (at LREC 2016)*. Portorož.

Francis Bond, Piek Vossen, John McCrae, and Christiane Fellbaum. 2016b. CILI: The collaborative interlingual index. In *Proceedings of the 8th Global Wordnet Conference (GWC 2016)*, pages 50–57.

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`.

Sabrina Cerini, Valentina Compagnoni, Alice Demontis, Maicol Formentelli, and G Gandini. 2007. Micro-wnop: A gold standard for the evaluation of automatically compiled lexical resources for opinion mining. *Language resources and linguistic theory: Typology, second language acquisition, English linguistics*, pages 200–210.

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*, 41(13):5984–5994.

Paul Ekman. 1992. An argument for basic emotions. *Cognition & Emotion*, 6(3):169–200.

Andrea Esuli and Fabrizio Sebastiani. 2006. SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining. In *Proceedings of 5th Conference on Language Resources and Evaluation LREC 2006*, pages 417–422.

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

Marek Maziarz, Maciej Piasecki, Ewa Rudnicka, Stan Szpakowicz, and Paweł Kędzia. 2016. plwordnet 3.0 – a comprehensive lexical-semantic resource. In Nicoletta Calzolari, Yuji Matsumoto, and Rashmi Prasad, editors, *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*, pages 2259–2268. ACL, ACL. URL `http://aclweb.org/anthology/C/C16/`.

Robert Plutchik. 1980. *Emotion: Theory, research, and experience*, volume Vol. 1. Theories of emotion. Academic, New York.

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*.

Philip J. Stone, Dexter C. Dunphy, Marshall S. Smith, and Daniel M. Ogilvie. 1966. *The General Inquirer: A Computer Approach to Content Analysis*. MIT Press.

Carlo Strapparava and Alessandro Valitutti. 2004. WordNet-Affect: An affective extension of WordNet. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, pages 1083–1086.

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.

Yoshimitsu Torii, Dipankar Das, Sivaji Bandyopadhyay, and Manabu Okumura. 2011. A Developing Japanese WordNet Affect for Analyzing Emotions. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2011), 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT 2011)*, pages 80–86.

Peter D. Turney and Michael L. Littman. 2003. Measuring Praise and Criticism: Inference of Semantic Orientation from Association. *ACM Transactions on Information Systems*, 21(4):315–346.

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.

Monika Zaśko-Zielińska, Maciej Piasecki, and Stan Szpakowicz. 2015. A large wordnet-based sentiment lexicon for Polish. In Ruslan Mitkov, Galia Angelova, and Kalina Boncheva, editors, *Proceedings of the International Conference Recent Advances in Natural Language Processing – RANLP'2015*, pages 721—730. INCOMA Ltd. Shoumen, BULGARIA, Hissar, Bulgaria. URL `http://aclweb.org/anthology/R15-1092`.