Almost-Unsupervised Cross-Language Opinion Analysis at NTCIR-7

Taras Zagibalov John Carroll
Department of Informatics
University of Sussex
Brighton BN1 9QH, UK
T.Zagibalov@sussex.ac.uk J.A.Carroll@sussex.ac.uk

Abstract

We describe the Sussex NLCL System entered in the NTCIR-7 Multilingual Opinion Analysis Task (MOAT). Our main focus is on the problem of portability of natural language processing systems across languages. Our system was the only one entered for all four of the MOAT languages, Japanese, English, and Simplified and Traditional Chinese. The system uses an almost-unsupervised approach applied to two of the sub-tasks: opinionated sentence detection and topic relevance detection.

Keywords: *NTCIR, opinion analysis, relevance detection, cross-language portability.*

1 Introduction

In our entry to the NTCIR-7 Multilingual Opinion Analysis Task (MOAT) [1] we focused on developing a system that can be easily ported from one language to another.

The most obvious way to design such a system is to make it as unsupervised as possible. Unsupervised systems derive all their information from raw, unannotated language data. Unsupervised techniques are promising for systems that need to be ported across domains, text types and languages. However, one of the main problems of unsupervised systems is that they are usually less accurate than supervised systems, which have access to annotated training data. Another problem is that unsupervised methods often need a large amount of raw data to be able derive any useful information.

As well as not using annotated data, unsupervised systems typically also contain few built-in assumptions about how a language is structured, encoded for example as rules or via a lexicon. Rule-based systems might consist of hundreds of rules and writing these manually may be as costly in time and linguistic expertise as annotating a corpus to be used by a supervised system. We therefore use no language-specific rules or other kinds of processing which would render our system less portable.

In this paper we present a system for opinion analysis which uses an unsupervised approach combined with a very limited amount of manual intervention.

We participated in two of the MOAT sub-tasks: opinionated sentence detection (required for all participants) and topic relevance detection (an optional sub-task). We applied our system to all four of the languages, Japanese, English, and Simplified and Traditional Chinese.

2 Our Approach

The relevance detection subsystem is completely unsupervised and finds relevant fragments of texts by comparing ranks of words (based on a word's relative frequency) in different topics. The underlying idea is that topic-relevant words have the biggest variance in relative frequency across topics.

For the opinion (subjectivity) classification sub-system, we automatically create a set of candidate opinion-bearing words from the training data, select a small number of these manually, and then extend this list with the words that are most strongly associated with them in the data.

2.1 Lexical Item Extraction

One of the most evident challenges of cross-lingual NLP is the different notion of 'word' in different languages. Among the MOAT languages, three out of four are in scripts that do not explicitly indicate word boundaries (Japanese, Simplified Chinese and Traditional Chinese). Trying as far as possible to avoid language-specific techniques, we did not use preliminary word segmentation in the Japanese or either of the Chinese experiments. For all languages we used the same routine for finding basic lexical units: split all texts at non-characters (punctuation, digits, etc., including any language-specific symbols) except space. All chunks were added to the list of lexical items. Then the lexical items were split at space (only relevant of course for English) and also added to the list together with their relative frequencies. The texts in Chinese and Japanese were split at characters, for example the Japanese string

５月に国連で開いた核拡散防止条約（ＮＰＴ）再検討会議を最後まで紛糾させた テーマもこの問題だった。

was split into

５月に国連で開いた核拡散防止条約（ＮＰＴ）再検討会議を最後まで紛糾させたテーマもこの問題だった。

This technique generates many possible sequences of lexical units which contain a lot of noise. One of the most frequent kinds of noise is when shorter sequences form parts of longer sequences. For example, if there is a sequence *ABC*, the method above also finds *A*, *B*, *C*, *AB* and *BC*. Some of these items may be valid words and phrases that are used independently from the larger unit they are part of, but some of them could never be used independently. To filter out parts of longer lexical units we compared their frequencies: if the frequency of a unit is the same as the frequency of a shorter sub-unit, the latter is deleted from the list. For example if *ABC* has frequency *Y*, *AB* has frequency *Y*, and *A* has frequency *X*, then only *ABC* and *A* are left in the list, while *AB* is deleted. After such filtering we used the 10% most frequent words for subsequent processing. This procedure was carried out for each topic, storing the resulting lists of lexical items separately for each topic.

2.2 Relevance Analysis

A decision about whether a sentence is related to a topic or not was made based on a list of lexical items which were regarded as topic relevance indicators. We derived this list as follows. For all lexical items from a given topic we calculated their relative frequency and assigned ranks to all of them (the most frequent being assigned the value 1, and so on). We then compared the ranks of words from a given topic to their ranks in any other topic:

$$V = |R_i - R_j|$$

If a word was not used in any other topic its value was therefore just its rank:

$$V = R_i$$

The average value of *V* was computed as

$$\bar{V} = \sum \frac{V}{n}$$

where *n* is the number of topics. Words with the largest average value of *V* were taken as the indicators.

For example, for the topic 'What is the relationship between AOL and Netscape?' the indicators were:

> *america online, appliances, designed, dominant, link, maker, netscape, online, services, start-ups, sun, technological change, they have, windows operating*

From manual inspection the words in this list look quite relevant, although there is some noise.

2.3 Opinionated Item Extraction

To determine whether a sentence is opinionated we used a semi-automatically generated list of words which are considered to be indicators of subjectivity. Knowing that such indicators are domain/topic dependent [2], we first tried to derive lists of words specific to each topic. However, poor results in initial experiments suggested that none of the topic-specific corpora for any of the four languages was large enough, so we merged all the topics together. The candidate list of opinion words was created as follows.

First, for each word in the list of most frequent words (see section 2.1) we found its immediate neighbours (words occurring either immediately before or after). Then for each word and neighbour we calculated the χ^2 score; neighbours for which $\chi^2 > 3.84$ were retained and sorted in decreasing order of χ^2 score, and the others discarded. Words having similar sets of neighbours might be semantically close. However, we want to avoid words that are related syntactically and not sematically, which we filtered out by considering first-order co-occurrence. For example, assume we have words *A*, *B* and *C*, each with neighbours as follows:

Word	Immediate context
A	*X Y Z*
B	*A Y Z*
C	*B Y Z*

The input corpus must have contained the string *AB* or *BA* (since *A* has been observed in the immediate context of *B*). Similarly, *BC* is also a first-order co-occurrence. On the other hand, *A* and *C* are probably related semantically rather than syntactically since there is no first-order co-occurrence and both appear in the context of *Y* and *Z*. So the pairs *AB* and *BC* are filtered out as syntactic, and *AC* remains as probably being semantic.

To estimate degree of semantic association, we calculated a score *S* between every remaining pair of words, measuring the similarity of neighbours lists:

$$S = \sum \frac{1}{r}$$

where the sum is over the neighbours present in both neighbours lists, and *r* is the rank of a neighbour in the list of the first word. The word pairs were then filtered to leave only the most associated ones. We used two filters. The first filtered out all pairs with *S* less than $\bar{x}$ -1.96σ. The second filter deleted all words that occurred unusually often (threshold $\bar{x}$ $+1.96\sigma$); such words are often function words without any task-relevant value. Finally we were left with a list of pairs of semantically highly associated words.

2.4 Opinionated Word Selection

From the list of pairs of associated words we selected those words which are relevant to the opinion task. Unfortunately we were not able to come up with an automatic technique of separating subjectivity markers from other words. Instead, we looked though the lists manually, selecting those words that looked most relevant to the task. In all, we spent less than one hour doing this for each language. Table 1 shows the lists of selected words, and Table 2 gives the numbers of words in the original and final lists. Neither of the authors knows any Japanese so we relied mostly on a dictionary when selecting Japanese words (although the first author's knowledge of Chinese characters helped a lot). If either of us had known Japanese we would undoubtedly have produced a better list. We also did not investigate which features are really relevant for subjectivity classification in any of the languages (for example, markers of modality, tense or aspect). Despite these two issues our system performed relatively well, indicating that the overall approach is viable.

After the list of subjectivity markers was derived it was applied to the corpus. If a sentence contained at least one of those words it was classified as opinionated. In the overall results, this system is called NLCL-1 (corresponding to 'Group=NLCL, RunID=1' in the official results).

The NLCL-1 system in general achieves high precision but low recall. In order to improve recall we tried two ways of expanding the list of manually-selected subjectivity markers. The first way included all words that were associated with the manually selected subjectivity markers (system NLCL-3). An alternative method included only those words whose association score was higher than the arithmetic mean for this list (system NLCL-2). As an example, the list for the English NLCL-2 system was:

> *active, advanced, analysts, common, developed, developing, difficult, easily, economists, effective, frequent, grave, hotel, immediate, important, likely, long, nino, notably, obvious, optimistic, played, popular, possess, primary, recently, robust, scientists, striking, successful, supervision, surprising, they will be, threaten, troubled, urgent, vital, vulnerable*

Table 1: Manually-selected opinionated words.

Chinese (Traditional)	Chinese (Simplified)	Japanese	English
難	太	難	*important*
功	比	激	*difficult*
害	最	貧	*effective*
感	强	悲	*popular*
好	欢	困難	*successful*
才	好	良	*easily*
最	良	可能	*troubled*
太	可能	戦闘	*striking*
利	善	深刻	*best*
效	害	焦点	*bad*
利用	难	犠牲	*painful*
認為	压力	強	*strong*
最	紧	最	*good*
	强	悪	
	恐	汚	

Table 2: Sizes of the lists of words.

Task	Automatically generated list	Number of selected words
Chinese (Traditional)	1154	13
Chinese (Simplified)	494	15
Japanese	491	15
English	1363	13

3 Evaluation Results

3.1 Traditional and Simplified Chinese

For the Chinese relevance and opinion sub-tasks (see Tables 3 and 4) our results are the lowest of all the systems, although not by much. More encouragingly, though, the results for each of our systems on the two sets of Chinese sub-tasks are numerically better than the results we obtained for the other two languages (see sections 3.2 and 3.3 below).

3.2 Japanese

We originally entered only a single system for the Japanese language sub-tasks, NLCL-1 (which uses just the manually selected list of 13 subjectivity markers). Table 5 shows the results. After the official submission, we also tested system NLCL-3 (which uses the manual list plus all associated words), to investigate whether the gains in recall would outweigh expected decreases

Table 3: Results for Chinese (Traditional). We erroneously submitted the results for system NLCL-2 as RunID=3, and NLCL-3 as RunID=2, so there is a minor disparity between this table and the official results.

	Sub-task	Precision (%)	Recall (%)	F-value
NLCL-1				
Lenient	Relevance	84.9	14.5	24.8
	Opinion	53.6	26.8	35.7
Strict	Relevance	92.4	18.0	30.1
	Opinion	62.6	29.3	39.9
NLCL-2				
Lenient	Relevance	86.4	28.6	43.0
	Opinion	49.4	50.6	50.0
Strict	Relevance	93.0	34.1	49.9
	Opinion	60.1	52.5	56.1
NLCL-3				
Lenient	Relevance	85.7	41.1	55.6
	Opinion	47.6	74.2	58.0
Strict	Relevance	92.8	48.5	63.7
	Opinion	58.3	74.1	65.3

Table 4: Results for Chinese (Simplified).

	Sub-task	Precision (%)	Recall (%)	F-value
NLCL-1				
Lenient	Relevance	96.3	32.6	48.7
	Opinion	44.3	39.9	42.0
Strict	Relevance	97.4	33.3	49.6
	Opinion	38.6	40.2	39.4
NLCL-2				
Lenient	Relevance	97.5	28.0	43.5
	Opinion	48.2	36.9	41.8
Strict	Relevance	98.5	28.5	44.1
	Opinion	44.3	39.0	41.4
NLCL-3				
Lenient	Relevance	97.1	58.5	73.0
	Opinion	43.2	69.9	53.4
Strict	Relevance	98.3	59.0	73.7
	Opinion	36.7	70.6	48.3

Table 5: Results for Japanese. *Note that the NLCL-3 results were obtained after the official submission.

	Sub-task	Precision (%)	Recall (%)	F-value
NLCL-1				
Lenient	Relevance	53.7	18.9	28.0
	Opinion	42.6	22.3	29.3
Strict	Relevance	30.1	21.1	24.8
	Opinion	31.4	22.6	26.3
*NLCL-3**				
Lenient	*Relevance*	*47.7*	*63.8*	*54.6*
	Opinion	*30.2*	*91.0*	*45.3*
Strict	*Relevance*	*22.7*	*61.1*	*33.1*
	Opinion	*22.2*	*91.9*	*35.8*

Table 6: Results for English.

	Sub-task	Precision (%)	Recall (%)	F-value
NLCL-1				
Lenient	Relevance	13.0	6.8	9.0
	Opinion	37.8	10.1	16.0
Strict	Relevance	5.3	8.5	6.5
	Opinion	11.7	10.5	11.1
NLCL-2				
Lenient	Relevance	17.5	14.4	15.8
	Opinion	33.8	18.6	24.0
Strict	Relevance	7.4	18.8	10.7
	Opinion	10.9	20.1	14.1
NLCL-3				
Lenient	Relevance	48.2	68.9	56.7
	Opinion	27.7	84.6	41.7
Strict	Relevance	16.4	72.7	26.8
	Opinion	8.4	86.1	15.3

in precision. This indeed turned out to be the case, with overall F-values at least 10 points better. The results for NLCL-1 are in the bottom quartile of systems, whereas those for NLCL-3 are in the third quartile.

3.3 English

For the English language tasks the NLCL-3 system performed well, delivering excellent results compared to other systems in the relevance sub-task, under both lenient and strict scoring (Table 6). In the opinion sub-task, NLCL-3 is in the third quartile.

The full set of official results for MOAT is presented in [1].

4 Related Work

There has been little previous work specifically directed at portability of opinion analysis systems. However, one could expect that unsupervised and semi-supervised approaches would be more portable than supervised systems.

Turney's SO-PMI approach [3] identifies words with positive and negative semantic orientation by measuring degree of association with seed words in a very large corpus, using the resulting set of words to classify documents with respect to overall sentiment direction. Baroni and Stefano [4] describe a similar technique for ranking a large list of adjectives according to subjectivity without resorting to any knowledge-intensive external resources (such as lexical databases, parsers or manual annotation). Zagibalov and Carroll [2] present a bootstrapping approach for finding sentiment-bearing words, which requires only a small corpus of opinionated documents and knowledge of frequent patterns of negation in the language. They demonstrate that their technique can find good sets of domain-relevant sentiment-bearing words with no manual intervention.

There has been some previous work addressing portability of information extraction systems across domains. For example, AutoSlog-TS [5] creates a dictionary of textual patterns that can be used to extract relevant facts from documents, using only a set of pre-classified documents as input, combined with a set of statistical tests and user-driven filtering and labelling of candidate extraction patterns. The approach was tested successfully on the MUC-4 terrorism domain. More recent work has applied similar approaches to sentence-level classification of subjectivity/objectivity. Riloff and Wiebe [6], apply a set of high-precision classifiers to a training corpus to learn extraction patterns; the newly found patterns are then used to find further subjective sentences from which to extract patterns in order to improve coverage. In more recent work, they use manually written rules to classify sentences as subjective or objective by looking for strong, well-established clues [7]. This step produces a high precision corpus of labelled sentences which is input to a supervised machine learning algorithm to create the final subjectivity classifier.

5 Conclusions and Future Work

We took part in the relevance and opinion sub-tasks for all four languages of the NTCIR-7 MOAT evaluation. The reason we participated in all languages was to evaluate the cross-lingual portability of our almost-unsupervised opinion analysis system.

Our results vary widely across the four languages: for the Japanese and English sub-tasks we obtained results which compare favourably with other systems, whereas in both Simplified and Traditional Chinese our system performed poorly in comparison with the other systems —although our results were numerically superior to those we obtained for the other languages. At this point we are not sure why our system's performance varies so much across the languages, and in particular why our system performed comparatively less well on the Chinese data. One possible explanation is that the Chinese data is more homogeneous and so more tractable for competing approaches based on supervised machine learning. Another possibility is that our system makes mistakes due to lack of language-specific processing (for example, by not using a word segmenter for Chinese). Nevertheless, our system was not far behind other systems even in the Chinese language tests, so it has achieved some success as a portability-oriented system.

In future we will attempt to analyse the factors which made our results so different across the languages. In Chinese and Japanese we will investigate the impact of using an automatic word segmentation module. We will also try other, language-independent techniques for finding words in Chinese and Japanese text.

Ultimately, we aim to devise techniques which would make our system completely unsupervised, and thus even more portable, by automating the filtering of opinionated words.

Acknowledgments

The first author is supported by the Ford Foundation International Fellowships Program.

References

[1] Y. Seki, D. Kirk Evans, L-W. Ku, L. Sun, H-H. Chen and N. Kando. Overview of Multilingual Opinion Analysis Task at NTCIR-7. In *Proceedings of the 7th NTCIR Evaluation Workshop*, Tokyo, Japan, 2008.

[2] T. Zagibalov and J. Carroll. Automatic seed word selection for unsupervised sentiment classification of Chinese text. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING)*, pages 1073–1080, Manchester, UK, 2008.

[3] P. D. Turney. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting of the Association for*

Computational Linguistics, pages 417–424, Philadelphia, Pennsylvania, 2002.

[4] M. Baroni and V. Stefano. Identifying subjective adjectives through web-based mutual information. In *Proceedings of the 7th KONVENS*, pages 17–24, Vienna, Austria, 2004.

[5] E. Riloff. Automatically generating extraction patterns from untagged text. In *Proceedings of the 13th National Conference on Artificial Intelligence (AAAI)*, pages 1044–1049, Portland, Oregon, 1996.

[6] E. Riloff and J. Wiebe. Learning extraction patterns for subjective expressions. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, pages 105–112, Sapporo, Japan, 2003.

[7] J. Wiebe and E. Riloff. Creating subjective and objective sentence classifiers from unannotated texts. In *Proceedings of the 6th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing)*, pages 475–486, Mexico City, 2005.