Evaluating Multilanguage-Comparability of Subjectivity Analysis Systems

Jungi Kim, Jin-Ji Li and Jong-Hyeok Lee
Division of Electrical and Computer Engineering
Pohang University of Science and Technology, Pohang, Republic of Korea
{yangpa,ljj,jhlee}@postech.ac.kr

Abstract

Subjectivity analysis is a rapidly growing field of study. Along with its applications to various NLP tasks, much work have put efforts into multilingual subjectivity learning from existing resources. Multilingual subjectivity analysis requires language-independent criteria for comparable outcomes across languages. This paper proposes to measure the multilanguage-comparability of subjectivity analysis tools, and provides meaningful comparisons of multilingual subjectivity analysis from various points of view.

1 Introduction

The field of NLP has seen a recent surge in the amount of research on subjectivity analysis. Along with its applications to various NLP tasks, there have been efforts made to extend the resources and tools created for the English language to other languages. These endeavors have been successful in constructing lexicons, annotated corpora, and tools for subjectivity analysis in multiple languages.

There are multilingual subjectivity analysis systems available that have been built to monitor and analyze various concerns and opinions on the Internet; among the better known are OASYS from the University of Maryland that analyzes opinions on topics from news article searches in multiple languages (Cesarano et al., 2007)[1] and TextMap, an entity search engine developed by Stony Brook University for sentiment analysis along with other functionalities (Bautin et al., 2008).[2] Though these systems currently rely on English analysis tools and a machine translation (MT) technology to translate other languages into English, up-to-date research provides various ways to analyze subjectivity in multilingual environments.

Given sentiment analysis systems in different languages, there are many situations when the analysis outcomes need to be multilanguage-comparable. For example, it has been common these days for the Internet users across the world to share their views and opinions on various topics including music, books, movies, and global affairs and incidents, and also multinational companies such as Apple and Samsung need to analyze customer feedbacks for their products and services from many countries in different languages. Governments may also be interested in monitoring terrorist web forums or its global reputation. Surveying these opinions and sentiments in various languages involves merging the analysis outcomes into a single database, thereby objectively comparing the result across languages.

If there exists an ideal subjectivity analysis system for each language, evaluating the multilanguage-comparability would be unnecessary because the analysis in each language would correctly identify the exact meanings of all input texts regardless of the language. However, this requirement is not fulfilled with current technology, thus the need for defining and measuring the multilanguage-comparability of subjectivity analysis systems is evident.

This paper proposes to evaluate the multilanguage-comparability of multilingual subjectivity analysis systems. We build a number of subjectivity classifiers that distinguishes subjective texts from objective ones, and measure the multilanguage-comparability according to our proposed evaluation method. Since subjectivity analysis tools in languages other than English are not readily available, we focus our experiments on comparing different methods to build multilingual analysis systems from the resources and systems

[1] http://oasys.umiacs.umd.edu/oasysnew/

[2] http://www.textmap.com/

Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages 595–603,
Uppsala, Sweden, 11-16 July 2010. ©2010 Association for Computational Linguistics

created for English. These approaches enable us to extend a monolingual system to many languages with a number of freely available NLP resources and tools.

2 Related Work

Much research have been put into developing methods for multilingual subjectivity analysis recently. With the high availability of subjectivity resources and tools in English, an easy and straightforward approach would be to employ a machine translation (MT) system to translate input texts in target languages into English then carry out the analyses using an existing subjectivity analysis tool (Kim and Hovy, 2006; Bautin et al., 2008; Banea et al., 2008). Mihalcea et al. (2007) and Banea et al. (2008) proposed a number of approaches exploiting a bilingual dictionary, a parallel corpus, and an MT system to port the resources and systems available in English to languages with limited resources.

For subjectivity lexicons translation, Mihalcea et al. (2007) and Wan (2008) used the first sense in a bilingual dictionary, Kim and Hovy (2006) used a parallel corpus and a word alignment tool to extract translation pairs, and Kim et al. (2009) used a dictionary to translate and a link analysis algorithm to refine the matching intensity.

To overcome the shortcomings of available resources and to take advantage of ensemble systems, Wan (2008) and Wan (2009) explored methods for developing a hybrid system for Chinese using English and Chinese sentiment analyzers. Abbasi et al. (2008) and Boiy and Moens (2009) have created manually annotated gold standards in target languages and studied various feature selection and learning techniques in machine learning approaches to analyze sentiments in multilingual web documents.

For learning multilingual subjectivity, the literature tentatively concludes that translating lexicon is less dependable in terms of preserving subjectivity than corpus translation (Mihalcea et al., 2007; Wan, 2008), and though corpus translation results in modest performance degradation, it provides a viable approach because no manual labor is required (Banea et al., 2008; Brooke et al., 2009).

Based on the observation that the performances of subjectivity analysis systems in comparable experimental settings for two languages differ, Banea et al. (2008) have attributed the variations in the difficulty level of subjectivity learning to the differences in language construction. Bautin et al. (2008)'s system analyzes the sentiment scores of entities in multilingual news and blogs and adjusted the sentiment scores using entity sentiment probabilities of languages.

Texts with an identical negative sentiment:
* The iPad could cannibalize the e-reader market.
* 아이패드가(iPad) 전자책 시장을(e-reader market) 위축시킬 수 있다(could cannibalize).

Texts with different strengths of positive sentiments:
* Samsung cell phones have excellent battery life.
* 삼성(Samsung) 휴대전화(cell phone) 배터리는 (battery) 그럭저럭(somehow or other) 오래간다(last long).

Figure 1: Examples of sentiments in multilingual text

3 Multilanguage-Comparability

3.1 Motivation

The quality of a subjectivity analysis tool is measured by its ability to distinguish subjectivity from objectivity and/or positive sentiments from negative sentiments. Additionally, a multilingual subjectivity analysis system is required to generate unbiased analysis results across languages; the system should base its outcome solely on the subjective meanings of input texts irrespective of the language, and the equalities and inequalities of subjectivity labels and intensities must be useful within and throughout the languages.

Let us consider two cases where the pairs of multilingual inputs in English and Korean have identical and different subjectivity meanings (Figure 1). The first pair of texts carry a negative sentiment about how the release of a new electronics device might affect an emerging business market. When a multilanguage-comparable system is inputted with such a pair, its output should appropriately reflect the negative sentiment, and be identical for both texts. The second pair of texts share a similar positive sentiment about a mobile device's battery capacity but with different strengths. A good multilingual system must be able to identify the positive sentiments and distinguish the differences in their intensities.

However, these kinds of conditions cannot be measured with performance evaluations independently carried out on each language; A system with a dissimilar ability to analyze subjective expressions from one language to another may deliver opposite labels or biased scores on texts with an identical subjective meaning, and vice versa, but still might produce similar performances on the evaluation data.

Macro evaluations on individual languages cannot provide any conclusions on the system's multilanguage-comparability capability. To measure how much of a system's judgment principles are preserved across languages, an evaluation from a different perspective is necessary.

3.2 Evaluation Approach

An evaluation of multilanguage-comparability may be done in two ways: measuring agreements in the outcomes of a pair of multilingual texts with an identical subjective meaning, or measuring the consistencies in the label and/or accordance in the order of intensity of a pair of texts with different subjectivities.

There are advantages and disadvantages to each approaches. The first approach requires multilingual texts aligned at the level of specificity, for instance, document, sentence and phrase, that the subjectivity analysis system works. Text corpora for MT evaluation such as newspapers, books, technical manuals, and government official records provide a wide variety of parallel texts, typically at the sentence level. Annotating these types of corpus can be efficient; as parallel texts must have identical semantic meanings, subjectivity–related annotations for one language can be projected into other languages without much loss of accuracy.

The latter approach accepts any pair of multilingual texts as long as they are annotated with labels and/or intensity. In this case, evaluating the label consistency of a multilingual system is only as difficult as evaluating that of a monolingual system; we can produce all possible pairs of texts from test corpora annotated with labels for each language. Evaluating with intensity is not easy for the latter approach; if test corpora already exist with intensity annotations for both languages, normalizing the intensity scores to a comparable scale is necessary (yet is uncertain unless every pair is checked manually), otherwise every pair of multilingual texts needs a manual annotation with its relative order of intensity.

In this paper, we utilize the first approach because it provides a more rational means; we can reasonably hypothesize that text translated into another language by a skilled translator carries an identical semantic meaning and thereby conveys identical subjectivity. Therefore the required resource is more easily attained in relatively inexpensive ways.

For evaluation, we measure the consistency in the subjectivity labels and the correlation of subjectivity intensity scores of parallel texts. Section 5.1 describes the details of evaluation metrics.

4 Multilingual Subjectivity System

We create a number of multilingual systems consisting of multiple subsystems each processing a language, where one system analyzes English, and the other systems analyze the Korean, Chinese, and Japanese languages. We try to reproduce a set of systems using diverse methods in order to compare the systems and find out which methods are more suitable for multilanguage-comparability.

4.1 Source Language System

We adopt the three systems described below as our source language systems: a state-of-the-art subjectivity classifier, a corpus-based, and a lexicon-based systems. The resources needed for developing the systems or the system itself are readily available for research purposes. In addition, these systems cover the general spectrum of current approaches to subjectivity analysis.

State-of-the-art (S-SA): OpinionFinder is a publicly-available NLP tool for subjectivity analysis (Wiebe and Riloff, 2005; Wilson et al., 2005).[3] The software and its resources have been widely used in the field of subjectivity analysis, and it has been the de facto standard system against which new systems are validated. We use a high-coverage classifier from the OpinionFinder's two sentence-level subjectivity classifiers. This Naive Bayes classifier builds upon a corpus annotated by a high-precision classifier with the bootstrapping of the corpus and extraction patterns. The classifier assesses a sentence's subjectivity with a label and a score for confidence in its judgment.

Corpus-based (S-CB): The MPQA opinion corpus is a collection of 535 newspaper articles in English annotated with opinions and private states at

[3] http://www.cs.pitt.edu/mpqa/opinionfinderrelease/, version 1.5

the sub-sentence level (Wiebe et al., 2003).[4] We retrieve the sentence level subjectivity labels for 11,111 sentences using the set of rules described in (Wiebe and Riloff, 2005). The corpus provides a relatively balanced corpus with 55% subjective sentences. We train an ML-based classifier using the corpus. Previous studies have found that, among several ML-based approaches, the SVM classifier generally performs well in many subjectivity analysis tasks (Pang et al., 2002; Banea et al., 2008).

We use SVM^{Light} with its default configurations,[5] inputted with a sentence represented as a feature vector of word unigrams and their counts in the sentence. An SVM score (a margin or the distance from a learned decision boundary) with a positive value predicts the input as being subjective, and negative value as objective.

Lexicon-based (S-LB): OpinionFinder contains a list of English subjectivity clue words with intensity labels (Wilson et al., 2005). The lexicon is compiled from several manually and automatically built resources and contains 6885 unique entries.

Riloff and Wiebe (2003) constructed a high-precision classifier for contiguous sentences using the number of strong and weak subjective words in current and nearby sentences. Unlike previous work, we do not (or rather, cannot) maintain assumptions about the proximity of input text. Using the lexicon, we build a simple and high-coverage rule-based subjectivity classifier. Setting the scores of strong and weak subjective words as 1.0 and 0.5, we evaluate the subjectivity of a given sentence as the sum of subjectivity scores; above a threshold, the input is subjective, and otherwise objective. The threshold value is optimized for an F-measure using the MPQA corpus, and is set to 1.0 throughout our experiments.

4.2 Target Language System

To construct a target language system leveraging on available resources in the source language, we consider three approaches from previous literature:

1. translating test sentences in target language into source language and inputting them into a source language system (Kim and Hovy, 2006; Bautin et al., 2008; Banea et al., 2008)
2. translating a source language training corpus into target language and creating a corpus-based system in target language (Banea et al., 2008)
3. translating a subjectivity lexicon from source language to target language and creating a lexicon-based system in target language (Mihalcea et al., 2007)

Each approach has its advantages and disadvantages. The advantage of the first approach is its simple architecture, clear separation of subjectivity and MT systems, and that it has only one subjectivity system, and is thus easier to maintain. Its disadvantage is that the time-consuming MT has to be executed for each text input. In the second and third approaches, a subjectivity system in the target language is constructed sharing corpora, rules, and/or features with the source language system. Later on, it may also include its own set of resources specifically engineered for the target language as a performance improvement. However, keeping the systems up-to-date would require as much effort as the number of languages. All three approaches use MT, and would suffer significantly if the translation results are poor.

Using the first approach, we can easily adopt all three source language systems;

- Target input translated into source, analyzed by source language system **S-SA**
- Target input translated into source, analyzed by source language system **S-CB**
- Target input translated into source, analyzed by source language system **S-LB**

The second and the third approaches are carried out as follows:

Corpus-based (T-CB): We translate the MPQA corpus into the target languages sentence by sentence using a web-based service.[6] Using the same method for **S-CB**, we train an SVM model for each language with the translated training corpora.

Lexicon-based (T-LB): This classifier is identical to **S-LB**, where the English lexicon is replaced by one of the target languages. We automatically translate the lexicon using free bilingual dictionaries.[7] First, the entries in the lexicon are looked

[4] http://www.cs.pitt.edu/mpqa/databaserelease/, version 1.2

[5] http://svmlight.joachims.org/, version 6.02

[6] Google Translate (http://translate.google.com/)

[7] quick_english-korean, quick_eng-zh_CN, and JMDict from StarDict (http://stardict.sourceforge.net/) licensed under GPL and EDRDG.

Table 1: Agreement on subjectivity (S for subjective, O objective) of 859 sentence chunks in Korean between two annotators (An. 1 and An. 2).

		An. 2		
		S	O	Total
An. 1	S	371	93	464
	O	23	372	395
	Total	394	465	859

Table 2: Agreement on projection of subjectivity (S for subjective, O objective) from Korean (KR) to English (EN) by one annotator.

		EN		
		S	O	Total
KR	S	458	6	464
	O	12	383	395
	Total	470	389	859

up in the dictionary, if they are found, we select the first word in the first sense of the definition. If the entry is not in the dictionary, we lemmatize it,[8] then repeat the search. Our simple approach produces moderate-sized lexicons (3,808, 3,980, 3,027 for Korean, Chinese, and Japanese) compared to Mihalcea et al. (2007)'s complicated translation approach (4,983 Romanian words). The threshold values are optimized using the MPQA corpus translated into each target language.[9]

5 Experiment

5.1 Experimental Setup

Test Corpus

Our evaluation corpus consists of 50 parallel newspaper articles from the Donga Daily News Website.[10] The website provides news articles in Korean and their human translations in English, Japanese, and Chinese. We selected articles that contain *Editorial* in its English title from a 30-day period. Three human annotators who are fluent in the two languages manually annotated N-to-N sentence alignments for each language pairs (KR-EN, KR-CH, KR-JP). By keeping only the sentence chunks whose Korean chunk appears in all language pairs, we were left with 859 sentence chunk pairs.

The corpus was preprocessed with NLP tools for each language,[11] and the Korean, Chinese, and Japanese texts were translated into English with the same web-based service used to translate the training corpus in Section 4.2.

Manual Annotation and Agreement Study

To assess the performance of our subjectivity analysis systems, the Korean sentence chunks were manually annotated by two native speakers of Korean with *Subjective* and *Objective* labels (Table 1). A proportion agreement of 0.86 and a kappa value of 0.73 indicate a substantial agreement between the two annotators. We set aside 743 sentence chunks that both annotators agreed on for the automatic evaluation of subjectivity analysis systems, thereby removing the borderline cases, which are difficult even for humans to assess. The corresponding sentence chunks for other languages were extracted and tagged with labels equivalent to Korean chunks.

In addition, to verify how consistently the subjectivity of the original texts is projected to the translated, we carried out another manual annotation and agreement study with Korean and English sentence chunks (Table 2).

Note that our cross-lingual agreement study is similar to the one carried out by Mihalcea et al. (2007), where two annotators labeled the sentence subjectivity of a parallel text in different languages. They reported that, similarly to monolingual annotations, most cases of disagreements on annotations are due to the differences in the annotators' judgments on subjectivity, and the rest from subjective meanings lost in the translation process and figurative language such as irony.

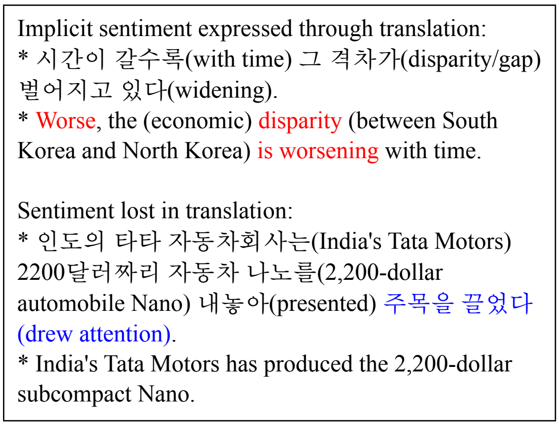
To avoid the role played by annotators' private views from disagreements, the subjectivity of sentence chunks in English were manually annotated by one of the annotators for the Korean text. Judged by the same annotator, we speculate that the disagreement in the annotation should account only for the inconsistency in the subjectivity projection. By proportion, the agreement between the annotation of Korean and English is 0.97, and the kappa is 0.96, suggesting an almost perfect agreement. Only a small number of sentence chunk pairs have inconsistent labels; six chunks in Ko-

[8] JWI (http://projects.csail.mit.edu/jwi/)

[9] Korean 1.0, Chinese 1.0, and Japanese 0.5

[10] http://www.donga.com/

[11] Stanford POS Tagger 1.5.1 and Stanford Chinese Word Segmenter 2008-05-21 (http://nlp.stanford.edu/software/), Chasen 2.4.4 (http://chasen-legacy.sourceforge.jp/), Korean Morphological Analyzer (KoMA) (http://kle.postech.ac.kr/)



Implicit sentiment expressed through translation:
* 시간이 갈수록(with time) 그 격차가(disparity/gap) 벌어지고 있다(widening).
* Worse, the (economic) disparity (between South Korea and North Korea) is worsening with time.

Sentiment lost in translation:
* 인도의 타타 자동차회사는(India's Tata Motors) 2200달러짜리 자동차 나노를(2,200-dollar automobile Nano) 내놓아(presented) 주목을 끌었다 (drew attention).
* India's Tata Motors has produced the 2,200-dollar subcompact Nano.

Figure 2: Excerpts from Donga Daily News with differing sentiments between parallel texts

rean lost subjectivity in translation, and implied subjective meanings in twelve chunks were expressed explicitly through interpretation. Excerpts from our corpus show two such cases (Figure 2).

Evaluation Metrics

To evaluate the multilanguage-comparability of subjectivity analysis systems, we measure 1) how consistently the system assigns subjectivity labels and 2) how closely numeric scores for systems' confidences correlate with regard to parallel texts in different languages.

In particular, we use Cohen's kappa coefficient for the first and Pearson's correlation coefficient for the latter. These widely used metrics provide useful comparability measures for categorical and quantitative data.

Both coefficients are scaled from -1 to $+1$, indicating negative to positive correlations. Kappa measures are corrected for chance, thereby yielding better measurements than agreement by proportion. The characteristics of Pearson's correlation coefficient that it measures linear relationships and is independent of change in origin, scale, and unit comply with our experiments.

5.2 Subjectivity Classification

Our multilingual subjectivity analysis systems were evaluated on the test corpora described in Section 5.1 (Table 3).

Due to the difference in testbeds, the performance of the state-of-the-art English system (**S-SA**) on our corpus is lower by about 10% relatively than the performance reported on the MPQA corpus.[12] However, it still performs sufficiently well and provides the most balanced results among the three source language systems; The corpus-based system (**S-CB**) classifies with a high precision, and the lexicon-based (**S-LB**) with a high recall. The source language systems (**S-SA,-CB,-LB**) lose a small percentage in precision when inputted with translations, but the recalls are generally on a par or even higher in the target languages.

For the systems created from target language resources, Corpus-based systems (**T-CB**) generally perform better than the ones with source language resource (**S-CB**), and lexicon-based systems (**T-LB**) perform worse than (**S-LB**). Similarly to systems with source language resources, **T-CB** classifies with a high precision and **T-LB** with a high recall, but the gap is less. Among the target languages, Korean tends to have a higher precision, and Japanese a higher recall than other languages in most systems.

Overall, **S-SA** provides easy accessibility when analyzing both the source and the target languages, with a balanced precision and recall performance. Among the other approaches, only **T-CB** is better in all measures than **S-SA**, and **S-LB** performs best on F-measure evaluations.

5.3 Multilanguage-Comparability

The evaluation results on multilanguage-comparability are presented in Table 4. The subjectivity analysis systems are evaluated with all language pairs with kappa and Pearson's correlation coefficients. Kappa and Pearson's correlation values are consistent with each other; Pearson's correlation between the two evaluation measures is 0.91.

We observe a distinct contrast in performances between corpus-based systems (**S-CB** and **T-CB**) and lexicon-based systems (**S-LB** and **T-LB**); All corpus-based systems show moderate agreements while agreements on lexicon-based systems are only fair.

Within corpus-based systems, **S-CB** performs better with language pairs that include English, and **T-CB** performs better with language pairs of the target languages.

For lexicon-based systems, systems in the target languages (**T-LB**) performs the worst with only slight to fair agreements between languages. Lexicon-based systems and state-of-the-art systems in the source language (**S-LB** and **S-SA**) result in average performances.

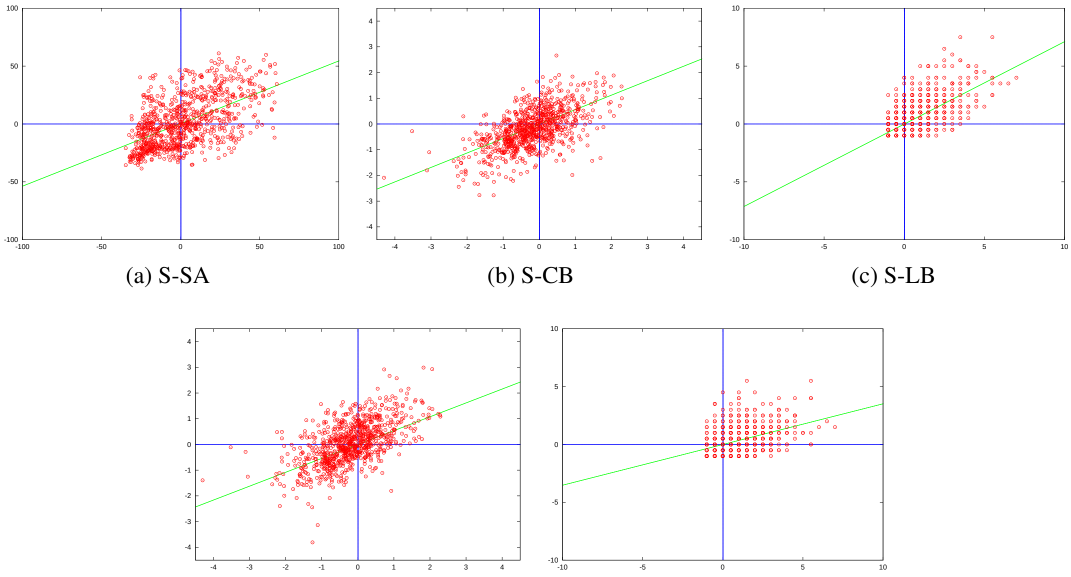
[12] precision, recall, and F-measure of 79.4, 70.6, and 74.7.

Table 3: Performance of subjectivity analysis with precision (P), recall (R), and F-measure (F). S-SA,-CB,-LB systems in Korean, Chinese, Japanese indicate English analysis systems inputted with translations of the target languages into English.

	English			Korean			Chinese			Japanese		
	P	R	F	P	R	F	P	R	F	P	R	F
S-SA	71.1	63.5	67.1	70.7	61.1	65.6	67.3	68.8	68.0	69.1	67.5	68.3
S-CB	**74.4**	53.9	62.5	**74.5**	52.2	61.4	71.1	63.3	67.0	**72.9**	65.3	68.9
S-LB	62.5	**87.7**	**73.0**	62.9	**87.7**	**73.3**	59.9	**91.5**	**72.4**	61.8	**94.1**	**74.6**
T-CB				72.4	67.5	69.8	**75.0**	66.2	70.3	72.5	70.3	71.4
T-LB				59.4	71.0	64.7	58.4	82.3	68.2	56.9	92.4	70.4

Table 4: Performance of multilanguage-comparability: kappa coefficient (κ) for measuring comparability of classification labels and Pearson's correlation coefficient (ρ) for classification scores for English (EN), Korean (KR), Chinese (CH), and Japanese (JP). Evaluations of T-CB,-LB for language pairs including English are carried out with results from S-CB,-LB for English and T-CB,-LB for target languages.

	S-SA		S-CB		S-LB		T-CB		T-LB	
	κ	ρ	κ	ρ	κ	ρ	κ	ρ	κ	ρ
EN & KR	0.41	0.55	**0.45**	**0.60**	0.37	0.59	0.42	**0.60**	0.25	0.41
EN & CH	0.39	0.54	**0.41**	**0.62**	0.33	0.52	0.39	0.57	0.22	0.38
EN & JP	0.39	0.53	**0.43**	**0.65**	0.30	0.59	0.40	0.59	0.15	0.33
KR & CH	0.36	0.54	0.39	0.59	0.28	0.57	**0.46**	**0.64**	0.23	0.37
KR & JP	0.37	0.60	0.44	0.69	0.50	0.69	**0.63**	**0.76**	0.18	0.38
CH & JP	0.37	0.53	**0.49**	**0.66**	0.29	0.57	0.46	0.63	0.22	0.46
Average	0.38	0.55	0.44	**0.64**	0.35	0.59	**0.46**	0.63	0.21	0.39



(a) S-SA

(b) S-CB

(c) S-LB

(d) T-CB

(e) T-LB

Figure 3: Scatter plots of English (x-axis) and Korean (y-axis) subjectivity scores from state-of-the-art (S-SA), corpus-based (S-CB), and lexicon-based (S-LB) systems of the source language, and corpus-based with translated corpora (T-CB), and lexicon-based with translated lexicon (T-LB) systems. Slanted lines in figures are best-fit lines through the origins.

Figure 3 shows scatter plots of subjectivity scores of our English and Korean test corpora evaluated on different systems; the data points on the first and the third quadrants are occurrences of label agreements, and the second and the fourth are disagreements. Linearly scattered data points are more correlated regardless of the slope.

Figure 3a shows a moderate correlation for multilingual results from the state-of-the-art system (**S-SA**). Agreements on objective instances are clustered together while agreements on subjective instances are diffused over a wide region.

Agreements between the source language corpus-based system (**S-CB**) and the corpus-based system trained with translated resources (**T-CB**) are more distinctively correlated than the results for other pairs of systems (Figures 3b and 3d). We notice that **S-CB** seems to have a lower number of outliers than **T-CB**, but slightly more diffusive.

Lexicon-based systems (**S-LB**, **T-LB**) generate noticeably uncorrelated scores (Figures 3c and 3e). We observe that the results from the English system with translated inputs (**S-LB**) is more correlated than those from systems with translated lexicons (**T-LB**), and that analysis results from both systems are biased toward subjective scores.

6 Discussion

Which approach is most suitable for multilingual subjectivity analysis?

In our experiments, the corpus-based systems trained on corpora translated from English to the target languages (**T-CB**) perform well for subjectivity classification and multilanguage-comparability measures on the whole. However, the methods we employed to expand the languages were naively carried out without much considerations for optimization. Further adjustments could improve the other systems for both classification and multilanguage-comparability performances.

Is there a correlation between classification performance and multilanguage-comparability?

Lexicon-based systems in the source language (**S-LB**) have good overall classification performances, especially on recall and F-measures. However, these systems performs worse on multilanguage-comparability than other systems with poorer classification performances. Intrigued by the observation, we tried to measure which criteria for classification performance influences multilanguage-comparability. We again employed Pearson's correlation metrics to measure the correlations of precision (P), recall (R), and F-measures (F) to kappa (κ) and Pearson's correlation (ρ) values.

Specifically, we measure the correlations between the sums of P, the sums of R, and the sums of F to κ and ρ for all pairs of systems.[13] The correlations of P with κ and ρ are 0.78 and 0.68, R -0.38 and -0.28, and F -0.20 and -0.05. These numbers strongly suggest that multilanguage-comparability correlates with the precisions of classifiers.

However, we cannot always expect a high-precision multilingual subjectivity classifier to be multilanguage-comparable as well. For example, the **S-SA** system has a much higher precision than **S-LB** consistently over all languages, but their multilanguage-comparability performances differed only by small amounts.

7 Conclusion

Multilanguage-comparability is an analysis system's ability to retain its decision criteria across different languages. We implemented a number of previously proposed approaches to learning multilingual subjectivity, and evaluated the systems on multilanguage-comparability as well as classification performance. Our experimental results provide meaningful comparisons of the multilingual subjectivity analysis systems across various aspects.

Also, we developed a multilingual subjectivity evaluation corpus from a parallel text, and studied inter-annotator, inter-language agreements on subjectivity, and observed persistent subjectivity projections from one language to another from a parallel text.

For future work, we aim extend this work to constructing a multilingual sentiment analysis system and evaluate it with multilingual datasets such as product reviews collected from different countries. We also plan to resolve the lexicon-based classifiers' classification bias towards subjective meanings with a list of objective words (Esuli and Sebastiani, 2006) and their multilingual expansion (Kim et al., 2009), and evaluate the multilanguage-comparability of systems constructed with resources from different sources.

[13] Pairs of values such as 71.1 + 70.7 and 0.41 for precisions and Kappa of S-SA for English and Korean.

Acknowledgement

We thank the anonymous reviewers for valuable comments and helpful suggestions. This work is supported in part by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (MEST) (2009-0075211), and in part by the BK 21 project in 2010.

References

Ahmed Abbasi, Hsinchun Chen, and Arab Salem. 2008. Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. *ACM Transactions on Information Systems*, 26(3):1–34.

Carmen Banea, Rada Mihalcea, Janyce Wiebe, and Samer Hassan. 2008. Multilingual subjectivity analysis using machine translation. In *EMNLP '08: Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 127–135, Morristown, NJ, USA.

Mikhail Bautin, Lohit Vijayarenu, and Steven Skiena. 2008. International sentiment analysis for news and blogs. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*.

Erik Boiy and Marie-Francine Moens. 2009. A machine learning approach to sentiment analysis in multlingual Web texts. *Information Retrieval*, 12:526–558.

Julian Brooke, Milan Tofiloski, and Maite Taboada. 2009. Cross-linguistic sentiment analysis: From english to spanish. In *Proceedings of RANLP 2009*, Borovets, Bulgaria.

Carmine Cesarano, Antonio Picariello, Diego Reforgiato, and V.S. Subrahmanian. 2007. The oasys 2.0 opinion analysis system: A demo. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*.

Andrea Esuli and Fabrizio Sebastiani. 2006. Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06)*, pages 417–422, Geneva, IT.

Soo-Min Kim and Eduard Hovy. 2006. Identifying and analyzing judgment opinions. In *Proceedings of the Human Language Technology Conference of the NAACL (HLT/NAACL'06)*, pages 200–207, New York, USA.

Jungi Kim, Hun-Young Jung, Sang-Hyob Nam, Yeha Lee, and Jong-Hyeok Lee. 2009. Found in translation: Conveying subjectivity of a lexicon of one language into another using a bilingual dictionary and a link analysis algorithm. In *ICCPOL '09: Proceedings of the 22nd International Conference on Computer Processing of Oriental Languages. Language Technology for the Knowledge-based Economy*, pages 112–121, Berlin, Heidelberg.

Rada Mihalcea, Carmen Banea, and Janyce Wiebe. 2007. Learning multilingual subjective language via cross-lingual projections. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL'07)*, pages 976–983, Prague, CZ.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 79–86.

Ellen Riloff and Janyce Wiebe. 2003. Learning extraction patterns for subjective expressions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Xiaojun Wan. 2008. Using bilingual knowledge and ensemble techniques for unsupervised Chinese sentiment analysis. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 553–561, Honolulu, Hawaii, October. Association for Computational Linguistics.

Xiaojun Wan. 2009. Co-training for cross-lingual sentiment classification. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 235–243, Suntec, Singapore, August. Association for Computational Linguistics.

Janyce Wiebe and Ellen Riloff. 2005. Creating subjective and objective sentence classifiers from unannotated texts. In *Proceedings of the 6th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2005)*, pages 486–497, Mexico City, Mexico.

Janyce Wiebe, E. Breck, Christopher Buckley, Claire Cardie, P. Davis, B. Fraser, Diane Litman, D. Pierce, Ellen Riloff, Theresa Wilson, D. Day, and Mark Maybury. 2003. Recognizing and organizing opinions expressed in the world press. In *Proceedings of the 2003 AAAI Spring Symposium on New Directions in Question Answering*.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT-EMNLP'05)*, pages 347–354, Vancouver, CA.