Multilingual Subjectivity: Are More Languages Better?

Carmen Banea, Rada Mihalcea
Department of Computer Science
University of North Texas
carmenbanea@my.unt.edu
rada@cs.unt.edu

Janyce Wiebe
Department of Computer Science
University of Pittsburgh
wiebe@cs.pitt.edu

Abstract

While subjectivity related research in other languages has increased, most of the work focuses on single languages. This paper explores the integration of features originating from multiple languages into a machine learning approach to subjectivity analysis, and aims to show that this enriched feature set provides for more effective modeling for the source as well as the target languages. We show not only that we are able to achieve over 75% macro accuracy in all of the six languages we experiment with, but also that by using features drawn from multiple languages we can construct high-precision meta-classifiers with a precision of over 83%.

1 Introduction

Following the terminology proposed by (Wiebe et al., 2005), subjectivity and sentiment analysis focuses on the automatic identification of private states, such as opinions, emotions, sentiments, evaluations, beliefs, and speculations in natural language. While subjectivity classification labels text as either subjective or objective, sentiment or polarity classification adds an additional level of granularity, by further classifying subjective text as either positive, negative or neutral.

To date, a large number of text processing applications have used techniques for automatic sentiment and subjectivity analysis, including automatic expressive text-to-speech synthesis (Alm et al., 1990), tracking sentiment timelines in on-line forums and news (Balog et al., 2006; Lloyd et al., 2005), and mining opinions from product reviews (Hu and Liu, 2004). In many natural language processing tasks, subjectivity and sentiment classification has been used as a first phase filtering to

generate more viable data. Research that benefited from this additional layering ranges from question answering (Yu and Hatzivassiloglou, 2003), to conversation summarization (Carenini et al., 2008), and text semantic analysis (Wiebe and Mihalcea, 2006; Esuli and Sebastiani, 2006a).

Although subjectivity tends to be preserved across languages – see the manual study in (Mihalcea et al., 2007), (Banea et al., 2008) hypothesize that subjectivity is expressed differently in various languages due to lexicalization, formal versus informal markers, etc. Based on this observation, our research seeks to answer the following questions. First, can we *reliably* predict sentence-level subjectivity in languages other than English, by leveraging on a manually annotated English dataset? Second, can we improve the English subjectivity classification by expanding the feature space through the use of multilingual data? Similarly, can we also improve the classifiers in the other target languages? Finally, third, can we benefit from the multilingual subjectivity space and build a high-precision subjectivity classifier that could be used to generate subjectivity datasets in the target languages?

The paper is organized as follows. We introduce the datasets and the general framework in Section 2. Sections 3, 4, and 5 address in turn each of the three research questions mentioned above. Section 6 describes related literature in the area of multilingual subjectivity. Finally, we draw our conclusions in Section 7.

2 Multilingual Datasets

Corpora that are manually annotated for subjectivity, polarity, or emotion, are available in only select languages, since they require a considerable amount of human effort. Due to this impediment, the focus of this paper is to create a method for extrapolating subjectivity data devel-

SubjP	SubjR	SubjF	ObjP	ObjR	ObjF	AllP	AllR	AllF
90.4%	34.2%	46.6%	82.4%	30.7%	44.7%	86.7%	32.6%	47.4%

Table 1: Results obtained with a rule-based subjectivity classifier on the MPQA corpus (Wiebe and Riloff, 2005)

oped in a source language and to transfer it to other languages. Multilingual feature spaces are generated to create even better subjectivity classifiers, outperforming those trained on the individual languages alone.

We use the Multi-Perspective Question Answering (MPQA) corpus, consisting of 535 English-language news articles from a variety of sources, manually annotated for subjectivity (Wiebe et al., 2005). Although the corpus is annotated at the clause and phrase levels, we use the sentence-level annotations associated with the dataset in (Wiebe and Riloff, 2005). A sentence is labeled as subjective if it has at least one private state of strength medium or higher. Otherwise the sentence is labeled as objective. From the approximately 9700 sentences in this corpus, 55% of them are labeled as subjective, while the rest are objective. Therefore, 55% represents the majority baseline on this corpus. (Wiebe and Riloff, 2005) apply both a subjective and an objective rule-based classifier to the MPQA corpus data and obtain the results presented in Table 1.[1]

In order to generate parallel corpora to MPQA in other languages, we rely on the method we proposed in (Banea et al., 2008). We experiment with five languages other than English (*En*), namely Arabic (*Ar*), French (*Fr*), German (*De*), Romanian (*Ro*) and Spanish (*Es*). Our choice of languages is motivated by several reasons. First, we wanted languages that are highly lexicalized and have clear word delimitations. Second, we were interested to cover languages that are similar to English as well as languages with a completely different etymology. Consideration was given to include Asian languages, such as Chinese or Japanese, but the fact that their script without word-segmentation preprocessing does not directly map to words was a deterrent. Finally, another limitation on our choice of languages is the need for a publicly available machine translation system between the source language and each of the target languages.

We construct a subjectivity annotated corpus for each of the five languages by using machine translation to transfer the source language data into the target language. We then project the original sentence level English subjectivity labeling onto the target data. For all languages, other than Romanian, we use the Google Translate service,[2] a publicly available machine translation engine based on statistical models. The reason Romanian is not included in this group is that, at the time when we performed the first experiments, Google Translate did not provide a translation service for this language. Instead, we used an alternative statistical translation system called LanguageWeaver,[3] which was commercially available, and which the company kindly allowed us to use for research purposes.

The raw corpora in the five target languages are available for download at http://lit.csci.unt.edu/index.php/Downloads, while the English MPQA corpus can be obtained from http://www.cs.pitt.edu/mpqa.

Given the specifics of each language, we employ several preprocessing techniques. For Romanian, French, English, German and Spanish, we remove all the diacritics, numbers and punctuation marks except - and '. The exceptions are motivated by the fact that they may mark contractions, such as En: *it's* or Ro: *s-ar* (*may be*), and the component words may not be resolved to the correct forms. For Arabic, although it has a different encoding, we wanted to make sure to treat it in a way similar to the languages with a Roman

---

[1]For the purpose of this paper we follow this abbreviation style: *Subj* stands for subjective, *Obj* stands for objective, and *All* represents overall macro measures, computed over the subjective and objective classes; *P*, *R*, *F*, and *MAcc* correspond to precision, recall, F-measure, and macro-accuracy, respectively.

---

[2]http://www.google.com/translate_t
[3]http://www.languageweaver.com/

alphabet. We therefore use a library[4] that maps Arabic script to a space of Roman-alphabet letters supplemented with punctuation marks so that they can allow for additional dimensionality.

Once the corpora are preprocessed, each sentence is defined by six views: one in the original source language (English), and five obtained through automatic translation in each of the target languages. Multiple datasets that cover all possible combinations of six languages taken one through six (a total of 63 combinations) are generated. These datasets feature a vector for each sentence present in MPQA (approximately 9700). The vector contains only unigram features in one language for a monolingual dataset. For a multilingual dataset, the vector represents a cumulation of monolingual unigram features extracted from each view of the sentence. For example, one of the combinations of six taken three is Arabic-German-English. For this combination, the vector is composed of unigram features extracted from each of the Arabic, German and English translations of the sentence.

We perform ten-fold cross validation and train Naïve Bayes classifiers with feature selection on each dataset combination. The top 20% of the features present in the training data are retained. For datasets resulting from combinations of all languages taken one, the classifiers are monolingual classifiers. All other classifiers are multilingual, and their feature space increases with each additional language added. Expanding the feature set by encompassing a group of languages enables us to provide an answer to two problems that can appear due to data sparseness. First, enough training data may not be available in the monolingual corpus alone in order to correctly infer labeling based on statistical measures. Second, features appearing in the monolingual test set may not be present in the training set and therefore their information cannot be used to generate a correct classification.

Both of these problems are further explained through the examples below, where we make the simplifying assumption that the words in italics are the only potential carriers of subjective content, and that, without them, their surrounding

contexts would be objective. Therefore, their association with an either objective or subjective meaning imparts to the entire segment the same labeling upon classification.

To explore the first sparseness problem, let us consider the following two examples extracted from the English version of the MPQA dataset, followed by their machine translations in German:

> "En 1: rights group Amnesty International said it was *concerned* about the high risk of violence in the aftermath"
> "En 2: official said that US diplomats to countries *concerned* are authorized to explain to these countries"
> "De 1: Amnesty International sagte, es sei *besorgt* über das hohe Risiko von Gewalt in der Folgezeit"
> "De 2: Beamte sagte, dass US-Diplomaten *betroffenen* Länder berechtigt sind, diese Länder zu erklären"

We focus our discussion on the word *concerned*, which in the first example is used in its subjective sense, while in the second it carries an objective meaning (as it refers to a group of countries exhibiting a particular feature defined earlier on in the context). The words in italics in the German contexts represent the translations of *concerned* into German, which are functionally different as they are shaped by their surrounding context. By training a classifier on the English examples alone, under the data sparseness paradigm, the machine learning model may not differentiate between the word's objective and subjective uses when predicting a label for the entire sentence. However, appending the German translation to the examples generates additional dimensions for this model and allows the classifier to potentially distinguish between the senses and provide the correct sentence label.

For the second problem, let us consider two other examples from the English MPQA and their respective translations into Romanian:

> "En 3: could secure concessions on Taiwan in return for *supporting* Bush on issues such as anti-terrorism and"

---

[4]Lingua::AR::Word PERL library.

Lang	SubjP	SubjR	SubjF	ObjP	ObjR	ObjF	AllP	AllR	AllF	MAcc
En	74.01%	**83.64%**	**78.53%**	**75.89%**	63.68%	**69.25%**	**74.95%**	**73.66%**	**73.89%**	**74.72%**
Ro	73.50%	82.06%	77.54%	74.08%	63.40%	68.33%	73.79%	72.73%	72.94%	73.72%
Es	**74.02%**	82.84%	78.19%	75.11%	**64.05%**	69.14%	74.57%	73.44%	73.66%	74.44%
Fr	73.83%	83.03%	78.16%	75.19%	63.61%	68.92%	74.51%	73.32%	73.54%	74.35%
De	73.26%	83.49%	78.04%	75.32%	62.30%	68.19%	74.29%	72.90%	73.12%	74.02%
Ar	71.98%	81.47%	76.43%	72.62%	60.78%	66.17%	72.30%	71.13%	71.30%	72.22%

Table 2: Naïve Bayes learners trained on six individual languages

"En 4: to the potential for change from within America. *Supporting* our schools and community centres is a good"

"Ro 3: ar putea asigura concesii cu privire la Taiwan, în schimb pentru *susţinerea* lui Bush pe probleme cum ar fi anti-terorismului şi"

"Ro 4: la potenţialul de schimbare din interiorul Americii. *Sprijinirea* şcolile noastre şi centre de comunitate este un bun"

In this case, *supporting* is used in both English examples in senses that are both subjective; the word is, however, translated into Romanian through two synonyms, namely *susţinerea* and *sprijinirea*. Let us assume that sufficient training examples are available to strengthen a link between *supporting* and *susţinerea*, and the classifier is presented with a context containing *sprijinirea*, unseen in the training data. A multilingual classifier may be able to predict a label for the context using the co-occurrence metrics based on *supporting* and extrapolate a label when the context contains both the English word and its translation into Romanian as *sprijinirea*. For a monolingual classifier, such an inference is not possible, and the feature is discarded. Therefore a multi-lingual classifier model may gain additional strength from co-occurring words across languages.

3 Question 1

Can we reliably predict sentence-level subjectivity in languages other than English, by leveraging on a manually annotated English dataset?

In (Banea et al., 2008), we explored several methods for porting subjectivity annotated data from a source language (English) to a target language (Romanian and Spanish). Here, we focus on the transfer of manually annotated corpora through the usage of machine translation by projecting the original sentence level annotations onto the generated parallel text in the target language. Our aim is not to improve on that method, but rather to verify that the results are reliable across a number of languages. Therefore, we conduct this experiment in several additional languages, namely French, German and Arabic, and compare the results with those obtained for Spanish and Romanian.

Table 2 shows the results obtained using Naïve Bayes classifiers trained in each language individually, with a macro accuracy ranging from 71.30% (for Arabic) to 73.89% (for English).[5] As expected, the English machine learner outperforms those trained on other languages, as the original language of the annotations is English. However, it is worth noting that all measures do not deviate by more than 3.27%, implying that classifiers built using this technique exhibit a consistent behavior across languages.

4 Question 2

Can we improve the English subjectivity classification by expanding the feature space through the use of multilingual data? Similarly, can we also improve the classifiers in the other target languages?

We now turn towards investigating the impact on subjectivity classification of an expanded feature space through the inclusion of multilingual data. In order to methodically assess classifier behavior, we generate multiple datasets containing all pos-

---

[5]Note that the experiments conducted in (Banea et al., 2008) were made on a different test set, and thus the results are not directly comparable across the two papers.

No lang	SubjP	SubjR	SubjF	ObjP	ObjR	ObjF	AllP	AllR	AllF
1	73.43%	82.76%	77.82%	74.70%	62.97%	68.33%	74.07%	72.86%	73.08%
2	74.59%	83.14%	78.63%	75.70%	64.97%	69.92%	75.15%	74.05%	74.28%
3	75.04%	83.27%	78.94%	76.06%	65.75%	70.53%	75.55%	74.51%	74.74%
4	75.26%	83.36%	79.10%	76.26%	66.10%	70.82%	75.76%	74.73%	74.96%
5	75.38%	83.45%	79.21%	76.41%	66.29%	70.99%	75.90%	74.87%	75.10%
6	**75.43%**	**83.66%**	**79.33%**	**76.64%**	**66.30%**	**71.10%**	**76.04%**	**74.98%**	**75.21%**

Table 3: Average measures for a particular number of languages in a combination (from one through six) for Naïve Bayes classifiers using a multilingual space

sible combinations of one through six languages, as described in Section 2. We then train Naïve Bayes learners on the multilingual data and average our results per each group comprised of a particular number of languages. For example, for one language, we have the six individual classifiers described in Section 3; for the group of three languages, the average is calculated over 20 possible combinations; and so on.

Table 3 shows the results of this experiment. We can see that the overall F-measure increases from 73.08% – which is the average over one language – to 75.21% when all languages are taken into consideration (8.6% error reduction). We measured the statistical significance of these results by considering on one side the predictions made by the best performing classifier for one language (i.e., English), and on the other side the predictions made by the classifier trained on the multilingual space composed of all six languages. Using a paired t-test, the improvement was found to be significant at $p = 0.001$. It is worth mentioning that both the subjective and the objective precision measures increase to 75% when more than 3 languages are considered, while the overall recall level stays constant at 74%.

To verify that the improvement is due indeed to the addition of multilingual features, and it is not a characteristic of the classifier, we also tested two other classifiers, namely KNN and Rocchio. Figure 1 shows the average macro-accuracies obtained with these classifiers. For all the classifiers, the accuracies of the multilingual combinations exhibit an increasing trend, as a larger number of languages is used to predict the subjectivity annotations. The Naïve Bayes algorithm has the best performance, and a relative error rate reduc-
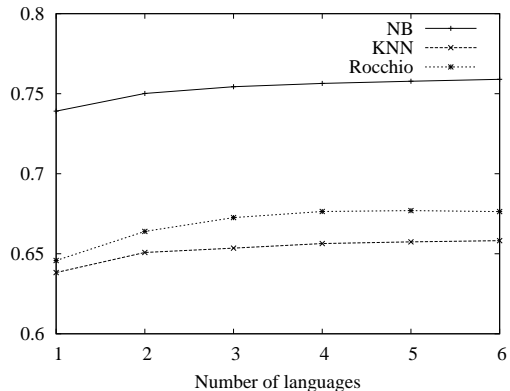

Figure 1: Average Macro-Accuracy per group of languages (combinations of 6 taken one through six)

tion in accuracy of 8.25% for the grouping formed of six languages versus one, while KNN and Rocchio exhibit an error rate reduction of 5.82% and 9.45%, respectively. All of these reductions are statistically significant.

In order to assess how the proposed multilingual expansion improves on the individual language classifiers, we select one language at a time to be the reference, and then compute the average accuracies of the Naïve Bayes learner across all the language groupings (from one through six) that contain the language. The results from this experiment are illustrated in Figure 2. The baseline in this case is represented by the accuracy obtained with a classifier trained on only one language (this corresponds to 1 on the X-axis). As more languages are added to the feature space, we notice a steady improvement in performance. When the language of reference is Arabic, we obtain an error reduction of 15.27%; 9.04% for Ro-
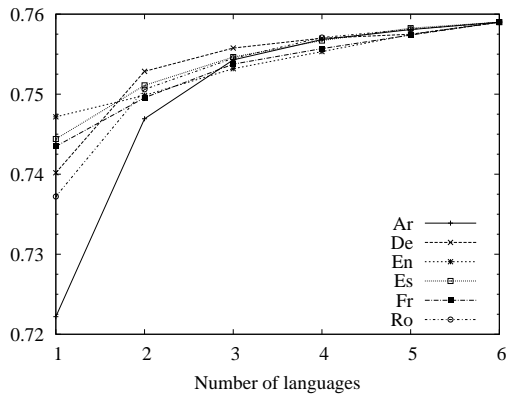
Figure 2: Average macro-accuracy progression relative to a given language

manian; 7.80% for German; 6.44% for French; 6.06% for Spanish; and 4.90 % for English. Even if the improvements seem minor, they are consistent, and the use of a multilingual feature set enables every language to reach a higher accuracy than individually attainable.

In terms of the best classifiers obtained for each grouping of one through six, English provides the best accuracy among individual classifiers (74.71%). When considering all possible combinations of six classifiers taken two, German and Spanish provide the best results, at 75.67%. Upon considering an additional language to the mix, the addition of Romanian to the German-Spanish classifier further improves the accuracy to 76.06%. Next, the addition of Arabic results in the best performing overall classifier, with an accuracy of 76.22%. Upon adding supplemental languages, such as English or French, no further improvements are obtained. We believe this is the case because German and Spanish are able to expand the dimensionality conferred by English alone, while at the same time generating a more orthogonal space. Incrementally, Romanian and Arabic are able to provide high quality features for the classification task. This behavior suggests that languages that are somewhat further apart are more useful for multilingual subjectivity classification than intermediary languages.

5 Question 3

Can we train a high precision classifier with a good recall level which could be used to generate subjectivity datasets in the target languages?

Since we showed that the inclusion of multilingual information improves the performance of subjectivity classifiers for all the languages involved, we further explore how the classifiers' predictions can be combined in order to generate high-precision subjectivity annotations. As shown in previous work, a high-precision classifier can be used to automatically generate subjectivity annotated data (Riloff and Wiebe, 2003). Additionally, the data annotated with a high-precision classifier can be used as a seed for bootstrapping methods, to further enrich each language individually.

We experiment with a majority vote meta-classifier, which combines the predictions of the *monolingual* Naïve Bayes classifiers described in Section 3. For a particular number of languages (one through six), all possible combinations of languages are considered. Each combination suggests a prediction only if its component classifiers agree, otherwise the system returns an "unknown" prediction. The averages are computed across all the combinations featuring the same number of languages, regardless of language identity.

The results are shown in Table 4. The macro precision and recall averaged across groups formed using a given number of languages are presented in Figure 3. If the average monolingual classifier has a precision of 74.07%, the precision increases as more languages are considered, with a maximum precision of 83.38% obtained when the predictions of all six languages are considered (56.02% error reduction). It is interesting to note that the highest precision meta-classifier for groups of two languages includes German, while for groups with more than three languages, both Arabic and German are always present in the top performing combinations. English only appears in the highest precision combination for one, five and six languages, indicating the fact that the predictions based on Arabic and German are more robust.

We further analyze the behavior of each language considering only those meta-classifiers that include the given language. As seen in Figure 4, all languages experience a boost in performance

No lang	SubjP	SubjR	SubjF	ObjP	ObjR	ObjF	AllP	AllR	AllF
1	73.43%	82.76%	77.82%	74.70%	62.97%	68.33%	74.07%	72.86%	73.08%
2	76.88%	76.39%	76.63%	80.17%	54.35%	64.76%	78.53%	65.37%	70.69%
3	78.56%	72.42%	75.36%	82.58%	49.69%	62.02%	80.57%	61.05%	68.69%
4	79.61%	69.50%	74.21%	84.07%	46.54%	59.89%	81.84%	58.02%	67.05%
5	80.36%	67.17%	73.17%	85.09%	44.19%	58.16%	82.73%	55.68%	65.67%
6	**80.94**%	65.20%	72.23%	**85.83**%	42.32%	56.69%	**83.38**%	53.76%	64.46%

Table 4: Average measures for a particular number of languages in a combination (from one through six) for meta-classifiers
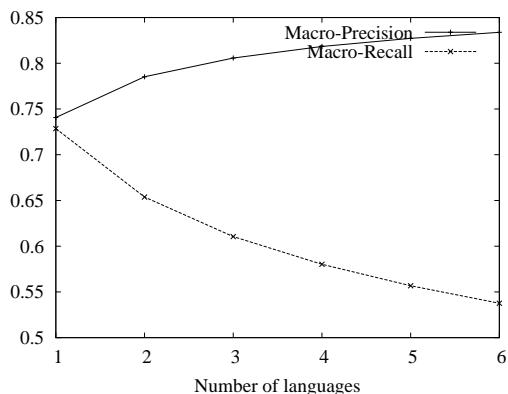

Figure 3: Average Macro-Precision and Recall across a given number of languages
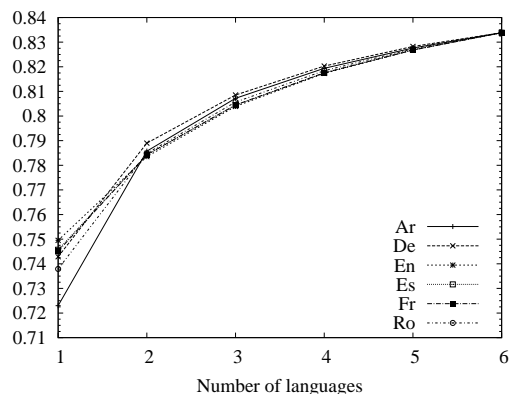

Figure 4: Average Macro-Precision relative to a given language

as a result of paired language reinforcement. Arabic gains an absolute 11.0% in average precision when considering votes from all languages, as compared to the 72.30% baseline consisting of the precision of the classifier using only monolingual features; this represents an error reduction in precision of 66.71%. The other languages experience a similar boost, including English which exhibits an error reduction of 50.75% compared to the baseline. Despite the fact that with each language that is added to the meta-classifier, the recall decreases, even when considering votes from all six languages, the recall is still reasonably high at 53.76%.

The results presented in table 4 are promising, as they are comparable to the ones obtained in previous work. Compared to (Wiebe et al., 2005), who used a high-precision rule-based classifier on the English MPQA corpus (see Table 1), our method has a precision smaller by 3.32%, but a recall larger by 21.16%. Additionally, unlike

(Wiebe et al., 2005), which requires language-specific rules, making it applicable only to English, our method can be used to construct a high-precision classifier in any language that can be connected to English via machine translation.

6 Related Work

Recently, resources and tools for sentiment analysis developed for English have been used as a starting point to build resources in other languages, via cross-lingual projections or monolingual and multi-lingual bootstrapping. Several directions were followed, focused on leveraging annotation schemes, lexica, corpora and automated annotation systems. The English annotation scheme developed by (Wiebe et al., 2005) for opinionated text lays the groundwork for the research carried out by (Esuli et al., 2008) when annotating expressions of private state in the Italian Content Annotation Bank. Sentiment and subjectivity lexica such as the one included with

the OpinionFinder distribution (Wiebe and Riloff, 2005), the General Inquirer (Stone et al., 1967), or the SentiWordNet (Esuli and Sebastiani, 2006b) were transfered into Chinese (Ku et al., 2006; Wu, 2008) and into Romanian (Mihalcea et al., 2007). English corpora manually annotated for subjectivity or sentiment such as MPQA (Wiebe et al., 2005), or the multi-domain sentiment classification corpus (Blitzer et al., 2007) were subjected to experiments in Spanish, Romanian, or Chinese upon automatic translation by (Banea et al., 2008; Wan, 2009). Furthermore, tools developed for English were used to determine sentiment or subjectivity labeling for a given target language by transferring the text to English and applying an English classifier on the resulting data. The labels were then transfered back into the target language (Bautin et al., 2008; Banea et al., 2008). These experiments are carried out in Arabic, Chinese, English, French, German, Italian, Japanese, Korean, Spanish, and Romanian.

The work closest to ours is the one proposed by (Wan, 2009), who constructs a polarity co-training system by using the multi-lingual views obtained through the automatic translation of product-reviews into Chinese and English. While this work proves that leveraging cross-lingual information improves sentiment analysis in Chinese over what could be achieved using monolingual resources alone, there are several major differences with respect to the approach we are proposing here. First, our training set is based solely on the automatic translation of the English corpus. We do not require an in-domain dataset available in the target language that would be needed for the co-training approach. Our method is therefore transferable to any language that has an English-to-target language translation engine. Further, we focus on using multi-lingual data from six languages to show that the results are reliable and replicable across each language and that multiple languages aid not only in conducting subjectivity research in the target language, but also in improving the accuracy in the source language as well. Finally, while (Wan, 2009) research focuses on polarity detection based on reviews, our work seeks to determine sentence-level subjectivity from raw text.

7   Conclusion

Our results suggest that including multilingual information when modeling subjectivity can not only extrapolate current resources available for English into other languages, but can also improve subjectivity classification in the source language itself. We showed that we can improve an English classifier by using out-of-language features, thus achieving a 4.90% error reduction in accuracy with respect to using English alone. Moreover, we also showed that languages other than English can achieve an F-measure in subjectivity annotation of over 75%, without using any manually crafted resources for these languages. Furthermore, by combining the predictions made by monolingual classifiers using a majority vote learner, we are able to generate sentence-level subjectivity annotated data with a precision of 83% and a recall level above 50%. Such high-precision classifiers may be later used not only to create subjectivity-annotated data in the target language, but also to generate the seeds needed to sustain a language-specific bootstrapping.

To conclude and provide an answer to the question formulated in the title, more languages are better, as they are able to complement each other, and together they provide better classification results. When one language cannot provide sufficient information, another one can come to the rescue.

Acknowledgments

References

Alm, Cecilia Ovesdotter, Dan Roth, and Richard Sproat. 1990. Emotions from text: machine learning for text-based emotion prediction. *Intelligence*.

Balog, Krisztian, Gilad Mishne, and Maarten De Rijke. 2006. Why Are They Excited? Identifying and Explaining Spikes in Blog Mood Levels. In *Proceedings of the*

11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-2006), Trento, Italy.

Banea, Carmen, Rada Mihalcea, Janyce Wiebe, and Samer Hassan. 2008. Multilingual Subjectivity Analysis Using Machine Translation. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP-2008)*, pages 127–135, Honolulu.

Bautin, Mikhail, 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-2008)*, Seattle, Washington.

Blitzer, John, Mark Dredze, and Fernando Pereira. 2007. Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational (ACL-2007)*, pages 440–447, Prague, Czech Republic. Association for Computational Linguistics.

Carenini, Giuseppe, Raymond T Ng, and Xiaodong Zhou. 2008. Summarizing Emails with Conversational Cohesion and Subjectivity. In *Proceedings of the Association for Computational Linguistics: Human Language Technologies (ACL- HLT 2008)*, pages 353–361, Columbus, Ohio.

Esuli, Andrea and Fabrizio Sebastiani. 2006a. Determining Term Subjectivity and Term Orientation for Opinion Mining. In *Proceedings of the 11th Meeting of the European Chapter of the Association for Computational Linguistics (EACL-2006)*, volume 2, pages 193–200, Trento, Italy.

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

Esuli, Andrea, Fabrizio Sebastiani, and Ilaria C Urciuoli. 2008. Annotating Expressions of Opinion and Emotion in the Italian Content Annotation Bank. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC-2008)*, Marrakech, Morocco.

Hu, Minqing and Bing Liu. 2004. Mining and Summarizing Customer Reviews. In *Proceedings of ACM Conference on Knowledge Discovery and Data Mining (ACM-SIGKDD-2004)*, pages 168–177, Seattle, Washington.

Ku, Lun-wei, Yu-ting Liang, and Hsin-hsi Chen. 2006. Opinion Extraction, Summarization and Tracking in News and Blog Corpora. In *Proceedings of AAAI-2006 Spring Symposium on Computational Approaches to Analyzing Weblogs*, number 2001, Boston, Massachusetts.

Lloyd, Levon, Dimitrios Kechagias, and Steven Skiena, 2005. *Lydia : A System for Large-Scale News Analysis (Extended Abstract) News Analysis with Lydia*, pages 161–166. Springer, Berlin / Heidelberg.

Mihalcea, Rada, 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-2007)*, pages 976–983, Prague, Czech Republic.

Riloff, Ellen and Janyce Wiebe. 2003. Learning Extraction Patterns for Subjective Expressions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2003)*, pages 105–112, Sapporo, Japan.

Stone, Philip J, Marshall S Smith, Daniel M Ogilivie, and Dexter C Dumphy. 1967. *The General Inquirer: A Computer Approach to Content Analysis. /*. The MIT Press, 1st edition.

Wan, Xiaojun. 2009. Co-Training for Cross-Lingual Sentiment Classification. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL-IJCNLP 2009)*, Singapore.

Wiebe, Janyce and Rada Mihalcea. 2006. Word Sense and Subjectivity. In *Proceedings of the joint conference of the International Committee on Computational Linguistics and the Association for Computational Linguistics (COLING-ACL-2006)*, Sydney, Australia.

Wiebe, Janyce and Ellen Riloff. 2005. Creating Subjective and Objective Sentence Classifiers from Unannotated Texts. In *Proceeding of CICLing-05, International Conference on Intelligent Text Processing and Computational Linguistics*, pages 486–497, Mexico City, Mexico.

Wiebe, Janyce, Theresa Wilson, and Claire Cardie. 2005. Annotating Expressions of Opinions and Emotions in Language. *Language Resources and Evaluation*, 39(2-3):165–210.

Wu, Yejun. 2008. Classifying attitude by topic aspect for English and Chinese document collections.

Yu, Hong and Vasileios Hatzivassiloglou. 2003. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentence. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2003)*, pages 129–136, Sapporo, Japan.