A Bootstrapping Method for Building Subjectivity Lexicons for Languages with Scarce Resources

Carmen Banea*, **Rada Mihalcea***, **Janyce Wiebe**‡

*Department of Computer Science
University of North Texas
carmenb@unt.edu, rada@cs.unt.edu

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

Abstract

This paper introduces a method for creating a subjectivity lexicon for languages with scarce resources. The method is able to build a subjectivity lexicon by using a small seed set of subjective words, an online dictionary, and a small raw corpus, coupled with a bootstrapping process that ranks new candidate words based on a similarity measure. Experiments performed with a rule-based sentence level subjectivity classifier show an 18% absolute improvement in F-measure as compared to previously proposed semi-supervised methods.

1. Introduction

There is growing interest in the automatic extraction of opinions, emotions, and sentiments in text (*subjectivity*), to provide tools and support for various natural language processing applications. Most of the research to date has focused on English, which is mainly explained by the availability of resources for subjectivity analysis, such as lexicons and manually labeled corpora.

In this paper, we describe a bootstrapping method for the automatic generation of a large subjectivity lexicon starting with a few seeds. Unlike previously proposed methods for building subjectivity lexicons, which typically rely on advanced language processing tools such as syntactic parsers or information extraction tools, our method specifically targets the construction of lexicons for languages with scarce resources. The method requires only a small set of seeds, a basic dictionary, and a small raw corpus, which makes it appealing for the large number of languages that have only limited text processing resources developed to date. We focus our experiments on Romanian, but the method is applicable to any other language.

2. Related Work

Many subjectivity and sentiment analysis tools rely on manually or semi-automatically constructed lexicons (Yu and Hatzivassiloglou, 2003; Riloff and Wiebe, 2003; Kim and Hovy, 2006). The availability of such lexicons enables the construction of efficient rule-based subjectivity and sentiment classifiers that rely on the presence of lexicon entries in the text.

Most of the work to date on subjectivity lexicon construction has assumed advanced natural language processing tools such as syntactic parsers (Wiebe, 2000) or tools for information extraction (Riloff and Wiebe, 2003), or the availability of broad-coverage rich lexical resources such as WordNet (Esuli and Sebastiani, 2006a). However, such tools and resources are available only for a handful of languages, which limits the applicability of these approaches. Instead, in the method introduced in this paper, we try to minimize the resources required to build a subjectivity lexicon. Thus, the method is potentially applicable to a large number of the languages spoken worldwide.

Our approach relates most closely to the method proposed by (Turney, 2002) for the construction of lexicons annotated for polarity. His algorithm starts with a few positive and negative seeds, and then uses data from the Web together with a similarity method (pointwise mutual information) to automatically grow this seed list. Our approach differs from (Turney, 2002) in two important ways: first, we do not address the task of polarity lexicon construction, but instead we focus on the acquisition of subjectivity lexicons. Second, Turney assumes a very large corpus such as the terabyte corpus of English documents available on the Web, whereas we rely on fewer, smaller-scale resources, namely a basic dictionary and a small raw corpus.

The problem of distinguishing subjective versus objective instances has often proven to be more difficult than subsequent polarity classification, so improvements in subjectivity classification promise to positively impact sentiment classification. This is reported in studies of manual annotation of phrases (Takamura et al., 2006), recognizing contextual polarity of expressions (Wilson et al., 2005), and sentiment tagging of words and word senses (Andreevskaia and Bergler, 2006; Esuli and Sebastiani, 2006b).

Another closely related work is our own previously proposed method for leveraging on resources available for English to construct resources for a second language (Mihalcea et al., 2007). That method assumed the availability of a bridge between languages, such as a bilingual lexicon or a parallel corpus. Instead, in the method proposed here, we rely exclusively on language-specific resources, and do not make use of any such bilingual resources which may not always be available.

3. Bootstrapping

Our method is able to quickly acquire a large subjectivity lexicon by bootstrapping from a few manually selected seeds. At each iteration, the seed set is expanded with related words found in an online dictionary, which are filtered by using a measure of word similarity. The bootstrapping process is illustrated in Figure 1.

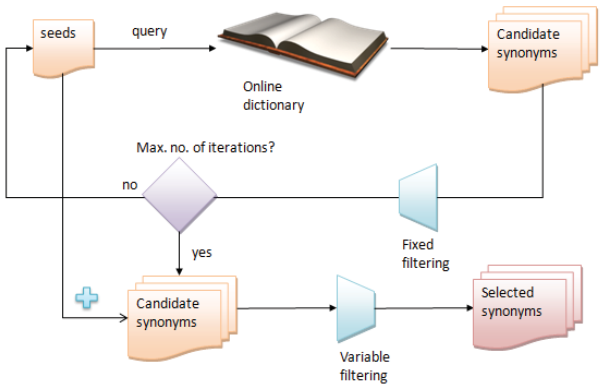

Figure 1: Bootstrapping process

3.1. Seed Set

We use a seed set of 60 seeds, evenhandedly sampled from verbs, nouns, adjectives and adverbs. The seeds were manually selected from two resources: the XI-th grade curriculum for Romanian Language and Literature developed by the Romanian Ministry of Education, which exemplifies students exerting proper use of subjective language, and translations of instances appearing in the OpinionFinder strong subjective lexicon (Wiebe and Riloff, 2005). Table 1 shows a sample of the entries in the initial seed set. A similar seed set can be easily obtained for any other language, either by finding a short listing of subjective words in the language of interest or by manually translating a small set of subjective entries from English.

Category	Sample entries (with their English translations)
Noun	blestem (curse), despot (tyrant), furie (fury), idiot (idiot), fericire (happiness)
Verb	iubi (love), aprecia (appreciate), spera (hope), dori (wish), uri (hate)
Adj	frumos (beautiful), dulce (sweet), urat (ugly), fericit (happy), fascinant (fascinating)
Adv	posibil (possibly), probabil (probably), desigur (of course), enervant (unnerving)

Table 1: Sample entries from the initial seed set

3.2. Dictionary

Starting with the seed set, new related words are added based on the entries found in a dictionary. For each seed word, we collect all the open-class words appearing in its definition, as well as synonyms and antonyms if available. Note that word ambiguity is not an issue, as the expansion is done with all the possible meanings for each candidate word, which are subsequently filtered for incorrect meanings by using the measure of word similarity.

In our experiments, we use an online Romanian dictionary http://www.dexonline.ro. Similar dictionaries are available for many languages; when online dictionaries are not available, they can be obtained with relatively low cost through OCR recognition performed on a hardcopy dictionary.

3.3. Bootstrapping Iterations

For each seed word, a query is formulated against the online Romanian dictionary. From the definitions obtained in this way, a list of related words is extracted, and added to the list of candidates if their own definition is part of the dictionary and if they do not appear in a list of stopwords. We then filter the candidate words based on their similarity with the original seed (see the following section), and continue to the next iteration until a maximum number of iterations is reached.

Note that the part-of-speech information is not maintained throughout the bootstrapping process, as words in the definitions belong to different parts-of-speech. Although the initial seed set is balanced with respect to syntactic categories, depending on the usage of words in definitions, this balance may be skewed toward one of the categories by the end of the bootstrapping process.

3.4. Filtering

In order to remove noise from the lexicon, we implemented a filtering step which is performed by calculating a measure of similarity between the original seeds and each of the possible candidates. We experimented with two corpus-based measures of similarity, namely the Pointwise Mutual Information (Turney, 2001) and Latent Semantic Analysis (LSA) (Dumais et al., 1988). We ultimately decided to use only LSA, as both methods provided similar results, but the LSA method was significantly faster and required less training data. After each iteration, only candidates with an LSA score higher than 0.4 (deduced empirically) between the original seed set and the candidates are considered to be expanded in the next iteration.

Upon bootstrapping termination, the subjectivity lexicons constructed incrementally after each iteration consist of a ranked list of candidates in decreasing order of similarity to the original seed set. A variable filtering threshold can be used to enforce the selection of only the most closely related candidates, resulting in more restrictive and pure subjectivity lexicons. In our experiments, we used the following thresholds: 0.40 (i.e. the lexicon resulting after the bootstrapping process without additional filtering), 0.50, 0.55, and 0.60.

The LSA module was trained on a half-million word Romanian corpus, consisting of a manually translated version of the SemCor balanced corpus (Miller et al., 1993). Corpora of similar size can be easily obtained for many low-resource languages by using semi-automatic methods for corpus construction (Ghani et al., 2001).

4. Evaluation and Discussion

For the evaluations, we use a subjectivity lexicon obtained through several iterations of bootstrapping, with an LSA similarity threshold of 0.5. Our experiments suggest that five bootstrapping iterations will be sufficient in extracting

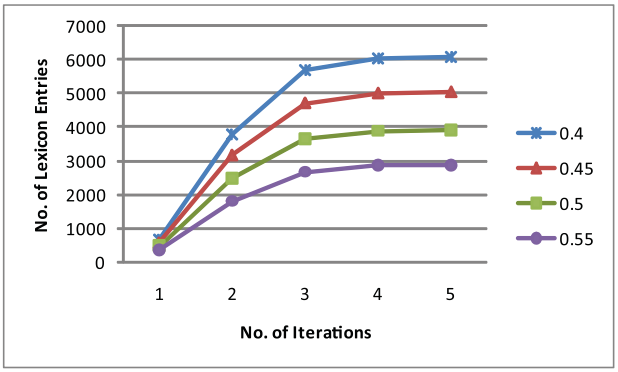

Figure 2: Lexicon Acquisition over 5 iterations

a subjectivity lexicon, as the number of features saturates during the last iteration (Figure 2). These settings resulted in a lexicon of 3,913 entries, which is used in a rule-based sentence-level subjectivity classifier. The classifier labels as subjective a sentence that contains three or more entries that appear in the subjective lexicon, and as objective a sentence that has two or fewer entries, respectively. This rule is derived based on the OpinionFinder rules (Wiebe and Riloff, 2005), which were modified to account for the fact that no strong/weak confidence labels are available.

We evaluate our results against a gold-standard corpus consisting of 504 Romanian sentences manually annotated for subjectivity. Two Romanian native speakers annotated the sentences individually, and the differences were adjudicated through discussions. The agreement of the two annotators is 0.83% (κ = 0.67); when the uncertain annotations are removed, the agreement rises to 0.89 (κ = 0.77). The two annotators reached consensus on all sentences for which they disagreed, resulting in a gold standard dataset with 272 (54%) subjective sentences and 232 (46%) objective sentences. More details about this data set are available in (Mihalcea et al., 2007).

The sentence-level subjectivity classification results are shown in Table 2. By using the extracted lexicon alone, we were able to obtain a rule-based subjectivity classifier with an overall F-measure of 61.69%.

To examine the effect of the number of bootstrapping iterations and the value of the LSA similarity threshold over the classifier, Table 2 displays the measures obtained through five bootstrapping iterations at an LSA threshold of 0.50, while Table 3 focuses on the fifth iteration tested over an LSA similarity of 0.40, 0.45, 0.50, 0.55, and 0.60. As expected, the overall F-measure is directly proportional to the LSA similarity score until the threshold becomes too restrictive, explicitly limiting the number of entries in the subjectivity lexicon.

We compare our results with those obtained by a previously proposed method that was based on a similar rule-based classifier. In (Mihalcea et al., 2007), a subjectivity lexicon was automatically obtained through the translation of the English subjectivity lexicon available in OpinionFinder. That lexicon consisted of 2,282 entries with a confidence label of *strong*, *neutral* or *weak* as flagged by the OpinionFinder lexicon. Table 4 shows the results obtained when using the translated lexicon to classify the subjectivity of

Iter.	Eval	Overall	Subj.	Obj.
1	P	57.56%	72.80%	52.60%
	R	57.56%	33.33%	85.59%
	F	57.56%	45.73%	65.16%
2	P	62.08%	64.93%	58.92%
	R	62.08%	63.74%	60.17%
	F	62.08%	64.33%	59.54%
3	P	61.30%	62.58%	59.42%
	R	61.30%	69.23%	52.12%
	F	61.30%	65.74%	55.53%
4	P	61.69%	62.83%	60.00%
	R	61.69%	69.96%	52.12%
	F	61.69%	66.20%	55.78%
5	P	61.69%	62.83%	60.00%
	R	61.69%	69.96%	52.12%
	F	61.69%	66.20%	55.78%

Table 2: Precision (P), Recall (R) and F-measure (F) for the bootstrapping subjectivity lexicon over 5 iterations and an LSA threshold of 0.5

LSA	Eval	Overall	Subj.	Obj.
0.40	P	60.12%	58.62%	66.02%
	R	60.12%	87.18%	28.81%
	F	60.12%	70.10%	40.12%
0.45	P	61.69%	60.60%	64.54%
	R	61.69%	81.69%	38.56%
	F	61.69%	69.58%	48.28%
0.50	P	61.69%	62.83%	60.00%
	R	61.69%	69.96%	52.12%
	F	61.69%	66.20%	55.78%
0.55	P	62.28%	68.49%	57.59%
	R	62.28%	54.95%	70.76%
	F	62.28%	60.98%	63.50%
0.60	P	54.81%	72.16%	50.73%
	R	54.81%	25.64%	88.56%
	F	54.81%	37.84%	64.51%

Table 3: Precision (P), Recall (R) and F-measure (F) for the 5th bootstrapping iteration for varying LSA scores

sentences in the same data set as used in our experiments. By comparing the results in Tables 2 and 4, we observe an absolute significant improvement of 18.03% in the overall F-measure when using the bootstrapping method introduced in this paper, as compared to the translated lexicon. Note that (Mihalcea et al., 2007) also proposed a corpus-based method for subjectivity classification; however that method is supervised and thus not directly comparable with the approach introduced in this paper. Interestingly, the F-measure obtained for the classification of subjective sentences is more than double in the case of the bootstrapping method, reflecting the ability of our approach to identify reliable subjective clues.

5. Conclusion

In this paper, we introduced a bootstrapping method able to quickly generate a large subjectivity lexicon that can be used to build rule-based sentence-level subjectivity classifiers for languages with scarce resources. The process starts with a small seed set of hand-picked subjective words, and with the help of an online dictionary, produces a lexicon of potential subjective candidates. The candidates are then

Eval	Overall	Subj.	Obj.
P	62.59%	80.00%	56.50%
R	33.53%	20.51%	48.91%
F	43.66%	32.65%	52.43%

Table 4: Precision (P), Recall (R) and F-measure (F) for the automatic translation subjectivity lexicon (2282 entries, cf. (Mihalcea et al., 2007))

ranked based on the LSA similarity measure, and the top approximately 4,000 entries are used to build a rule-based subjectivity classifier. Testing is performed between a human sentence-level annotated gold-standard and a heuristic providing sentence level automatic annotations. Even if unsupervised, our system is able to achieve a subjectivity F-measure of 66.20% and an overall F-measure of 61.69%. This system proposes a possible path towards identifying subjectivity in low-resource languages. In the future, we plan to experiment with variations of the bootstrapping mechanism, as well as with other similarity measures.

6. References

Alina Andreevskaia and Sabine Bergler. 2006. Mining wordnet for a fuzzy sentiment: Sentiment tag extraction from wordnet glosses. In *Proceedings of the 11rd Conference of the European Chapter of the Association for Computational Linguistics (EACL-2006)*.

S. T. Dumais, G. W. Furnas, T. K. Landauer, S. Deerwester, and R. Harshman. 1988. Using latent semantic analysis to improve access to textual information. In *CHI '88: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 281–285, New York, NY, USA. ACM.

A. Esuli and F. Sebastiani. 2006a. SentiWordNet: A publicly available lexical resource for opinion mining. In *Proceedings of the 5th Conference on Language Resources and Evaluation (LREC 2006)*, Genova, IT.

Andrea Esuli and Fabrizio Sebastiani. 2006b. Determining term subjectivity and term orientation for opinion mining. In *11th Meeting of the European Chapter of the Association for Computational Linguistics (EACL 2006)*.

R. Ghani, R. Jones, and D. Mladenic. 2001. Using the Web to create minority language corpora. In *Proceedings of the 10th International Conference on Information and Knowledge Management*.

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

R. Mihalcea, C. Banea, and J. Wiebe. 2007. Learning multilingual subjective language via cross-lingual projections. In *Proceedings of the Association for Computational Linguistics*, Prague, Czech Republic.

G. Miller, C. Leacock, T. Randee, and R. Bunker. 1993. A semantic concordance. In *Proceedings of the 3rd DARPA Workshop on Human Language Technology*, Plainsboro, New Jersey.

E. Riloff and J. Wiebe. 2003. Learning extraction patterns for subjective expressions. In *Conference on Empirical Methods in Natural Language Processing (EMNLP-03)*, pages 105–112.

Hiroya Takamura, Takashi Inui, and Manabu Okumura. 2006. Latent variable models for semantic orientations of phrases. In *Proceedings of the 11th Meeting of the European Chapter of the Association for Computational Linguistics (EACL 2006)*, Trento, Italy.

P. Turney. 2001. Mining the web for synonyms: PMI-IR versus LSA on TOEFL. In *Proceedings of the Twelfth European Conference on Machine Learning (ECML-2001)*, Freiburg, Germany.

P. Turney. 2002. 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 (ACL 2002)*, pages 417–424, Philadelphia.

J. Wiebe and E. 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) (invited paper)*, Mexico City, Mexico.

J. Wiebe. 2000. Learning subjective adjectives from corpora. In *Proceedings of the American Association for Artificial Intelligence (AAAI 2000)*, pages 735–740, Austin, Texas.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *hltemnlp2005*, pages 347–354, Vancouver, Canada.

Hong Yu and Vasileios Hatzivassiloglou. 2003. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Conference on Empirical Methods in Natural Language Processing (EMNLP-03)*, pages 129–136, Sapporo, Japan.