A Cross-Lingual Approach for Building Multilingual Sentiment Lexicons

Behzad Naderalvojoud[1][0000-0003-4429-5341], Behrang Qasemizadeh[2], Laura Kallmeyer[2], and Ebru Akcapinar Sezer[1]

[1] Hacettepe University, 06800 Beytepe, Ankara, Turkey
{n.behzad,ebru}@hacettepe.edu.tr
[2] DFG SFB 991
Universität Düsseldorf, Düsseldorf, Germany
{zadeh,kallmeyer}@phil.hhu.de

Abstract. We propose a cross-lingual distributional model to build sentiment lexicons in many languages from resources available in English. We evaluate our proposed method for two languages, German and Turkish, and several tasks. We show that the sentiment lexicons built using our method remarkably improve the performance of a lexicon-based BiLSTM sentiment classification model.

1 Introduction

Sentiment lexicons are important language resources for sentiment classification systems. The manual construction of these lexicons, however, is resource-intensive and thus expensive. When sentiment lexicons are not available for a language, one solution is to build them using automatic translation from available resources in other languages [19] such as the English *SentiWordNet* lexicon [1]. To this end, we propose a new cross-lingual distributional model to create a mapping between a pair of source–target languages so that the sentiment information about lexical items already known in the source language can be transferred to the target language.

We propose an extrinsic evaluation method to show the effectiveness of our method. We apply a stat-of-the-art neural-network *lexicon-based* sentiment classification method to a number evaluation datasets in German and Turkish using off-the-shelf sentiment lexicons. We then augment/replace these sentiment lexicons with lexicons that are built using our method and redo the sentiment classification tasks. We interpret the gain in the performance of the sentiment classifier in these tasks as the quality of our constructed lexicons, thus the effectiveness of our method.

In the remainder of this paper, Section 2 describes related work. Section 3 details our method. We report results from our experiments in Section 4 and conclude in Section 5.

2 Related Work

The dominant approach in sentiment analysis is to specify and use sentiment polarity (or the so-called subjectivity/objectivity) lexical databases, in which a word or phrase is often assigned to one or more quantities to describes its connotation (e.g., negative or positive) out-of-context. E.g., [11, 10] assume that the polarity of words is defined independently of their domain of usage and they assign a *general* polarity to each word. In this respect, most lexicons available for sentiment classification are built using a method similar to [9] and [21].

Enlarging sentiment lexicons built through a manual annotation effort is a popular topic in sentiment analysis. For example, a sentiment dictionary is built simply by assigning positive and negative sentiment values to a small set of seed words; and then, it is expanded using semantic relations that are available in other lexical databases such as WordNet. Similarly, distributional similarities can be used. An example is [17], in which two sets of seed words are collected manually and then expanded by finding words that are *most similar* to them using statistical measures such as Pointwise Mutual Information. [8] combines these ideas and expand the lists of positive and negative seed words using semantic relations asserted in WordNet and predicts the polarity of unseen words using two probabilistic models. A similar idea can be found in [5]. Provided that a resource such as WordNet is available, [1] shows that it is possible to build high quality sentiment lexicon such as SentiWordNet using automatic methods. Simply put, SentiWordNet (SWN) assigns polarity values to the WordNet synsets. However, machine-readable lexical knowledge bases such as WordNet are not available for many languages, and except for English, their coverage is often limited. Hence, these methods are not applicable to several languages, e.g. Turkish.

At the absence of high quality lexical knowledge bases, some studies have attempted to translate English resources such as SentiWordNet to other languages using machine translation techniques. E.g., [3] has generated sentiment lexicons from SentiWordNet for three Indian languages ('Bengali', 'Hindi' and 'Telugu') using a *word-level synset transfer technique*. Two sentiment lexicons have been proposed by [19] for German using a semi-automatic translation method from the *Subjectivity Clue List* [20] and *SentiSpin* [13]. Similarly, [18] compares methods for translating subjective terms in SentiWordNet to Turkish.

Last but not least, although a few studies (e.g., [1]) take into account word senses and assign more than one polarity to terms (i.e., depending on the inventory of senses for them), most work focuses on assigning polarities based on term usages in context and provide contextualized or domain-specific sentiment lexicons [7, 6].

3 Cross-Lingual Method for Building Sentiment Lexicon

We use English SWN (i.e., a sentiment lexicon organized around synsets) as input to our method. To use polarity values assigned to English synsets in a target language other than English, we must create a mapping between the target language words and WordNet's synsets. Hence, we build a model in which meanings of words in the target language are represented by WordNet synsets. Subsequently, we use this model to extract polarity values for words in the target language. Steps for deriving this model for a target language are described below:

Building a Cross-Lingual Distributional Model First, we generate a co-occurrence matrix from a sentences-aligned parallel corpus. From the input corpus, we extract a vocabulary $S = \{w_1 \ldots w_n\}$ for the source language and another one $T = \{w'_1 \ldots w'_m\}$ for the target language. We instantiate a matrix $\mathbf{M}_{n \times m}$ and use it to keep track of the counts of w_is and w'_js that co-occur in the aligned sentences. Note that S contains both words and multiword expressions of maximum length of 4 tokens. For the source language, we distinguish between words of different part-of-speech category (limited to nouns, verbs, adjectives and adverbs), e.g. instead of simply asserting the word-form *book* in S, we assert two entries *book-n* (i.e., the word book with the part-of-speech category noun) and *book-v* (with the part-of-speech category verb).

The obtained co-occurrence counts in matrix $\mathbf{M}$ are smoothed using a log-entropy transformations (similar to the one proposed in [16]). Each component m_{ij} of $\mathbf{M}$ is weighted using $m_{ij} = w_j \log(m_{ij} + 1)$, in which $w_j = 1 - \frac{H_j}{\log(n)}$ and H_j is the entropy of the column j of $\mathbf{M}$. That is, $H_j = -\sum_{i=1}^{n} p_{ij} \log(p_{ij})$, in which $p_{i,j} = \frac{m_{ij}}{\sum_{i=1}^{n} x_{ij}}$.

Synset Representation The weighted $\mathbf{M}$ is used to represent the subjective synsets of SWN. In SWN, the subjectivity of each synset is shown using three sentiment score p (positive polarity), n (negative polarity), and u (neutrality) for which $p+n+u = 1$. We assign a single subjectivity s value to each synset by subtracting the negative polarity score from the positive one (i.e. $s = p-n$). The sign of s indicates the overall sentiment of its synset (i.e., positive or negative). A synset in WordNet can be interpreted and understood using (a) its *gloss* which is a textual description that describe the meaning of the synset, and/or (b) by looking at the *synset terms*, i.e., the collection of terms/words that share the same meaning represented by the synset. Here, we exploit the latter. Accordingly, each synset x is represented by one vector $\boldsymbol{x}$; $\boldsymbol{x}$ is the sum of the row vectors in $\mathbf{M}$ that represent the terms that belong to x. We call these $\boldsymbol{x}$s synset vectors. We replacing row vectors of $\mathbf{M}$ with these synset vectors to form a synset-based co-occurance matrix $\mathbf{M}'_{|\boldsymbol{x}| \times m}$, where $|\boldsymbol{x}|$ is the number of synset vectors.

Synset Mapping In this step, we build a mapping between the target language words and synsets. Each target word j is mapped to k synsets: for target word j $(1 \leq j \leq m)$, we sort $m'_{ij} \in M'$ $(1 \leq i \leq |\boldsymbol{x}|)$ in descending order and choose top k synsets that are given by the index i. The polarities assigned to these top k synsets are set as the polarity values for the target word. This is the major difference of our method and the previous translation-based method for building sentiment lexicon: instead of using a word-by-word translation, we use a synset-to-word translation strategy which allows a target word to express several meanings of different sentiment polarities.

4 Evaluation and Empirical Experiments

To assess the effectiveness of our method and in order to show its impact on sentiment classification tasks, we report results from a number of empirical investigations. To build our distributional model, we use Open-Subtitle corpora [15], a set of sentence-aligned parallel corpora built from movie subtitles. As mentioned earlier, our source language is English. As target language we choose Turkish and German and report result based on models for the pairs of English-German and English-Turkish; details regarding the construction of these models and respectively the lexicon induced from them are given in Section 4.1.

To evaluate our method for building sentiment lexicons, we employ a lexicon-based deep learning method based on BiLSTM proposed in [14] for sentiment classification. In this approach, the sentiment score of a sentence is computed based on the weighted sum (an interpolation) of the polarity values of the subjective words obtained from the lexicon. Simply put, these weights are learned from training samples to modify the prior polarity values of words with respect to their usage context.

We conduct our experiments on two datasets (German twitter data (SB10K) [2] and German customer feedback (GermEval2017) [22]) for German and three datasets (hotel, movie, and product reviews) for Turkish.

German twitter data (SB10K) [2] consists of 9949 tweets that are labelled as *Positive*, *Negative* and *Neutral*. The original train and test sets are used in the experiment; we choose randomly 10% of the train data and use it as the development set. German customer feedback (GermEval2017) [22], which is used in the GermEval 2017 shared task, is accompanied by two types of test sets (synchronic and diachronic). For GermEval2017 dataset, as an additional baseline, we report the best-obtained result (*Best-GermEval*) from the shared task [22]. For Turkish, we use hotel review dataset proposed in [18] with its original split for train and test. The movie and product review datasets are proposed in [4]; we use (80% : 10% : 10%) splits as train, dev and test sets, respectively. In all the Turkish datasets, documents are labelled either as *Positive* or *Negative* class.

4.1 Lexicons

We created German and Turkish lexicons using the method proposed in Section 3 from roughly two million aliened sentences in OPUS. In our experiments, we choose $k = 10$ (i.e., target words are mapped to 10 synsets). Since each target word has more than one polarity score (based on the synset mappings), we propose four ways to produce a single polarity: (1) we use the average of all polarity scores (*avg*), (2) we sum all the scores (*sum*), (3) the score is obtained by calculating the percentage of the assigned positive and negative polarities to the word and the polarity with the majority of votes is used as the polarity of the word (*major*) and (4) the score is obtained by subtracting the percentage of the negative synsets from the percentage of the positive ones (*subMajor*). Note that (1) and (2) are calculated based on the polarity scores, whereas (3) and (4) are obtained by counting the number of positive or negative synsets.[3]

To build baselines, we repeat sentiment classification tasks using sentiment lexicons other than ones built by our method. For German, we employ the German sentiment lexicon proposed in [19] which uses the translation of English subjectivity clues [20]. The translation results from three online English-to-German translation systems were used to construct this German lexicon. It is worth noting that two other German lexicons are also available [19], however we selected the German subjectivity clue lexicon since the polarity values in this lexicon are assigned manually. For Turkish, we employ the sentiment lexicon proposed in [18]. This lexicon is built using a word-by-word translation of the subjective terms of SWN. Because this lexicon has been built using a 'parallel' translation method, we call it *parallel* in our experiments.

4.2 Result

GermEval2017 test sets: Tables 1 and 2 show the obtained results for the two test sets of the GermEval dataset. For the synchronic test set (Tables 1), without using a sentiment lexicon, our BiLSTM classifier yields a weak F-measure for both *Positive* and *Negative* classes. However, using a sentiment lexicon results improve noticeably; despite the lack of positive and negative instances in the train set (6% and 26%, respectively), the lexicon-based BiLSTM model achieves better results than the standard BiLSTM. Namely, BiLSTM achieves F-measure values of 24.11 and 65.56 by using *sum* and *major* lexicons, respectively. Moreover, we observe that lexicons built using our method outperform the model that uses the German subjectivity clue (subj.clue-BiLSTM). Similarly, we outperform the best system of the GermEval2017 share task. Both *sum* and *major* lexicons yield high macro and micro F-measure values.

Table 1. Results on GermEval2017 (synchronic test set)

Lexicon-Model	Pos-F1	Neg-F1	Neu-F1	Macro-F1	Micro-F1
BiLSTM	00.00	23.60	80.20	34.60	68.30
Best-GermEval	-	-	-	48.06	74.94
subj.clue-BiLSTM	12.33	62.77	81.65	52.25	74.20
avg-BiLSTM	13.63	63.51	82.32	53.15	75.13
major-BiLSTM	14.91	**65.56**	**82.68**	54.38	**75.72**
majorSub-BiLSTM	14.04	65.23	81.51	53.59	74.39
sum-BiLSTM	**24.11**	63.99	82.03	**56.71**	75.10

We observe similar results for the diachronic test set of the GermEval dataset. From Table 2, sum-BiLSTM gives the best result and it achieves the best macro

[3] The constructed lexicons are available at https://github.com/nbehzad/CLSL.

and micro F-measure values of 58.35 and 74.21, respectively. Similar to the synchronic test, we observe the positive effect of using lexicon-based sentiment data during classification. In this test, however, the subj.clue-BiLSTM model (i.e., our baseline) does not perform as well as it does in the synchronic test.

Table 2. Results on GermEval2017 (diachronic test set)

Lexicon-Model	Pos-F1	Neg-F1	Neu-F1	Macro-F1	Micro-F1
BiLSTM	00.00	25.20	81.60	35.60	70.00
Best-GermEval	-	-	-	51.65	73.62
subj.clue-BiLSTM	1.75	57.08	82.29	47.04	73.18
avg-BiLSTM	6.45	59.32	82.05	49.27	73.34
major-BiLSTM	26.67	59.07	81.83	55.86	73.13
majorSub-BiLSTM	15.07	58.86	82.16	52.03	73.18
sum-BiLSTM	**32.14**	**60.18**	**82.75**	**58.35**	**74.21**

SB10K test set: Table 3 shows the result. We observe that the proposed German sentiment lexicons, likewise previous tests, yield the best results; particularly, the sum method yields the best macro and micro F-measure values of 63.59 and 71.63, respectively.

Table 3. Results on the test set of SB10K dataset

Lexicon-Model	Pos-F1	Neg-F1	Neu-F1	Macro-F1	Micro-F1
BiLSTM	46.70	23.10	77.50	49.10	66.20
subj.clue-BiLSTM	62.90	42.99	79.34	61.74	69.91
avg-BiLSTM	63.74	39.30	79.01	60.68	69.82
major-BiLSTM	62.15	39.62	76.04	59.27	67.59
majorSub-BiLSTM	62.87	42.94	77.44	61.09	68.31
sum-BiLSTM	**66.23**	**44.52**	**80.03**	**63.59**	**71.63**

Turkish hotel and movie reviews: All the Turkish datasets have a balanced distribution of positive and negative instances, hence we report results only using micro F-measure and F-measures for the positive and negative classes. Table 4 reports the obtained results. We observe that using our method, the micro F-measure value increases from 0.7800 to 0.9007 in hotel reviews, and from 0.8450 to 0.8931 in movie reviews. Although, the sum-BiLSTM again produces more consistent results than the other methods, all the lexicon-based models perform better than the standard BiLSTM (as well as when using *parallel* lexicon) in both hotel and movie reviews.

Table 4. Results on Turkish hotel and movie reviews

Lexicon-Model	Hotel Review			Movie Review		
	Pos-F1	Neg-F1	Micro-F1	Pos-F1	Neg-F1	Micro-F1
BiLSTM	73.30	81.30	78.00	85.10	83.90	84.50
parallel-BiLSTM	79.97	85.42	83.12	87.51	87.90	87.71
avg-BiLSTM	85.79	88.60	87.34	88.44	89.21	88.84
major-BiLSTM	76.21	83.63	80.60	88.41	88.32	88.37
majorSub-BiLSTM	**89.35**	**90.70**	**90.07**	88.31	88.79	88.56
sum-BiLSTM	88.37	89.99	89.24	**89.06**	**89.54**	**89.31**

Turkish product reviews: To investigate the quality of Turkish sentiment lexicon built using our method in a cross-domain setting (e.g., as proposed in [12]), we repeat experiments over the product review dataset of 4 different domains (*books*, *DVD*, *electronics* and *kitchen* appliances). Table 5 reports the obtained results. As shown, all the sentiment lexicons consistently improve the performance of the base BiLSTM method in identifying negative reviews with an exception for the F-measure value for the positive class in the *kitchen* domain. However, the gain in performance using our method is higher than using the parallel lexicon.

Table 5. Results on Turkish product reviews

Lexicon-Model	Books		DVD		Electronics		Kitchen	
	Pos-F1	Neg-F1	Pos-F1	Neg-F1	Pos-F1	Neg-F1	Pos-F1	Neg-F1
BiLSTM	58.30	68.80	59.90	61.50	60.20	63.90	50.70	52.10
parallel-BiLSTM	**63.86**	73.29	63.70	66.20	79.69	81.63	37.25	64.04
avg-BiLSTM	62.90	70.51	**73.61**	72.06	85.71	**88.31**	44.04	64.33
major-BiLSTM	60.00	74.12	68.80	74.84	81.81	83.78	42.99	64.74
majorSub-BiLSTM	58.18	72.94	64.35	**75.15**	84.13	87.01	**58.33**	55.88
sum-BiLSTM	63.64	**76.47**	69.29	74.51	**86.57**	87.67	46.00	**70.00**

5 Conclusion

We proposed a cross-lingual approach for building sentiment lexicons in a target language from sentiment lexicons available in a source language. We showed the effectiveness of our proposed method and assessed the quality of the obtained lexicons through a number of evaluations. We improved results from a state-of-the-art lexicon-based BiLSTM sentiment classification system for German and Turkish in several tasks. The obtained results verified that lexicons generated by our proposed method can boost the performance of sentiment analysis and outperform other translation-based methods for building sentiment lexicons.

References

1. Baccianella, S., Esuli, A., Sebastiani, F.: SentiWordNet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In: LREC (2010)
2. Cieliebak, M., Deriu, J., Egger, D., Uzdilli, F.: A twitter corpus and benchmark resources for german sentiment analysis. SocialNLP 2017 p. 45 (2017)
3. Das, A., Bandyopadhyay, S.: SentiWordNet for Indian languages. In: ALR (2010)
4. Demirtas, E., Pechenizkiy, M.: Cross-lingual polarity detection with machine translation. In: WISDOM. p. 9. ACM (2013)
5. Esuli, A., Sebastiani, F.: Determining term subjectivity and term orientation for opinion mining. In: EACL. vol. 6, p. 2006 (2006)
6. Hung, C.: Word of mouth quality classification based on contextual sentiment lexicons. Information Processing & Management **53**(4), 751–763 (2017)
7. Hung, C., Chen, S.J.: Word sense disambiguation based sentiment lexicons for sentiment classification. Knowledge-Based Systems **110**, 224–232 (2016)
8. Kim, S.M., Hovy, E.: Determining the sentiment of opinions. In: COLING (2004)
9. Liu, B., Hu, M., Cheng, J.: Opinion observer: analyzing and comparing opinions on the web. In: WWW. pp. 342–351. ACM (2005)
10. Mohammad, S., Dunne, C., Dorr, B.: Generating high-coverage semantic orientation lexicons from overtly marked words and a thesaurus. In: EMNLP. pp. 599–608. ACL (2009)
11. Strapparava, C., Valitutti, A., Stock, O.: The affective weight of lexicon. In: LREC (2006)
12. Taboada, M., Brooke, J., Tofiloski, M., Voll, K., Stede, M.: Lexicon-based methods for sentiment analysis. Comput. Linguist. **37**(2), 267–307 (2011)
13. Takamura, H., Inui, T., Okumura, M.: Extracting semantic orientations of words using spin model. In: ACL. pp. 133–140. ACL (2005)
14. Teng, Z., Vo, D.T., Zhang, Y.: Context-sensitive lexicon features for neural sentiment analysis. In: EMNLP. pp. 1629–1638 (2016)
15. Tiedemann, J.: News from opus-a collection of multilingual parallel corpora with tools and interfaces. In: RNLP. vol. 5, pp. 237–248 (2009)
16. Turney, P.D.: Similarity of semantic relations. Comput. Linguist. **32**(3), 379–416 (2006)
17. Turney, P.D., Littman, M.L.: Measuring praise and criticism: Inference of semantic orientation from association. TOIS **21**(4), 315–346 (2003)
18. Ucan, A., Naderalvojoud, B., Sezer, E.A., Sever, H.: SentiWordNet for new language: Automatic translation approach. In: SITIS. pp. 308–315. IEEE (2016)
19. Waltinger, U.: GERMANPOLARITYCLUES: A lexical resource for German sentiment analysis. In: LREC. electronic proceedings, Valletta, Malta (May 2010)
20. Wilson, T., Wiebe, J., Hoffmann, P.: Recognizing contextual polarity in phrase-level sentiment analysis. In: EMNLP. pp. 347–354. ACL (2005)
21. Wilson, T., Wiebe, J., Hwa, R.: Just how mad are you? finding strong and weak opinion clauses. In: aaai. vol. 4, pp. 761–769 (2004)
22. Wojatzki, M., Ruppert, E., Holschneider, S., Zesch, T., Biemann, C.: Germeval 2017: Shared task on aspect-based sentiment in social media customer feedback. In: GermEval (2017)