Word Polarity Detection Using a Multilingual Approach

Cüneyd Murad Özsert and Arzucan Özgür

Department of Computer Engineering, Boğaziçi University,
Bebek, 34342 İstanbul, Turkey
muradozsert@gmail.com, arzucan.ozgur@boun.edu.tr

Abstract. Determining polarity of words is an important task in sentiment analysis with applications in several areas such as text categorization and review analysis. In this paper, we propose a multilingual approach for word polarity detection. We construct a word relatedness graph by using the relations in WordNet of a given language. We extend the graph by connecting the WordNets of different languages with the help of the Inter-Lingual-Index based on English WordNet. We develop a semi-automated procedure to produce a set of positive and negative seed words for foreign languages by using a set of English seed words. To identify the polarity of unlabeled words, we propose a method based on random walk model with commute time metric as proximity measure. We evaluate our multilingual approach for English and Turkish and show that it leads to improvement in performance for both languages.

Keywords: Semantic orientation, word polarity, sentiment analysis, random walk model, commute time, hitting time, WordNet.

1 Introduction

Identifying the semantic orientation or polarity of words is one of the most important topics in sentiment analysis. Many applications such as analyzing product/movie reviews (Morinaga et al., 2002; Turney, 2002; Popescu and Etzioni, 2005), and determining the attitudes of participants in online discussions (Hassan et al., 2010) are based on the polarities of the individual words.

Most previous studies on word polarity detection have been carried on for English and make use of language-specific resources such as WordNet (Miller, 1995) and General Inquirer (Stone et al., 1966). Wordnet, is a large lexical database for English, consisting of synsets (i.e. set of synonyms) each belonging to a distinct meaning. General Inquirer is an English lexicon, where words have been tagged with semantic categories such as positive and negative. In polarity detection studies WordNet has mainly been used to construct word relatedness graphs by connecting semantically related words and General Inquirer has been used to obtain labeled seed words for supervised settings and for evaluation purposes (Takamura et al., 2005; Hassan and Radev, 2010). Many languages do not have semantically tagged lexicons such as General Inquirer. Even though some of these languages have WordNets, they are in general not as comprehensive as the English WordNet. Most foreign WordNets such

A. Gelbukh (Ed.): CICLing 2013, Part II, LNCS 7817, pp. 75–82, 2013.

as EuroWordNet (Vossen, 1998) and BalkaNet (Tufiş et al., 2004) are structured in the same way as English WordNet (Miller, 1995) and are linked to each other with an Inter-Lingual-Index based on English WordNet.

In this work, we take advantage of the compatibility in WordNets and develop a multilingual approach for detecting polarities of English as well as foreign words. We construct a word-relatedness graph by not only connecting semantically related words in one WordNet but by also linking words from WordNets of different languages. We also propose a semi-automated method to generate labeled seed words for other languages by using the list of English seed words and the Inter-Lingual-Index. Then, we define a random walk over the word-relatedness graph from any given word to the set of positive and negative seed words. We use commute time as a proximity measure and classify a given word as positive if it is closer to the set of positive seed words compared to the negative seed words, and classify it as negative otherwise. We evaluate our approach for English and Turkish. Turkish WordNet (Bilgin et al., 2004) is completed within the BalkaNet project (Tufiş et al., 2004). It is constructed as being fully compatible with EuroWordNet, which in turn is compatible with English WordNet. We first show that our commute time model achieves performance comparable to the state-of-the-art in the literature. Then, we demonstrate that creating a multilingual word relatedness graph by connecting the WordNets of English and Turkish boosted the performance of word polarity detection for both languages. To our knowledge, we report the first results for Turkish word polarity detection and achieve an accuracy of 95%.

2 Related Work

Word polarity detection has been studied by several researchers in the past few years. Most of these studies have been evaluated for English words and are based on language resources available for English. For example, Turney and Littman (2003) propose an unsupervised algorithm, where they define seven positive and seven negative paradigm seed words. They use the English web corpus to query any given word with the paradigm words by using the near operator in a search engine. If the word tends to co-occur with positive paradigm words, it is classified as positive, and it is classified as negative otherwise. Takamura et al. (2005) propose a method, which regards semantic orientation as spin of electrons. They consider each word as an electron and its polarity as a spin value. They construct a word relatedness graph by using gloss definitions, thesaurus, and co-occurrence statistic for English. Words are classified as positive or negative according to their spin values. Hassan and Radev (2010) introduce a semi-supervised method where random walk model is used to find the polarities of English words. They construct a word relatedness graph by using the relations in English WordNet and use mean hitting time for polarity estimation.

Hassan et al. (2011) propose an algorithm to find semantic orientation of foreign words and evaluate their approach for Arabic and Hindi with a set of 300 manually labeled seed words for each language. They use random walk model with hitting time for polarity detection. They construct a multilingual network by connecting English and foreign words by using a Foreign-English dictionary. For every foreign word, they look up its possible meanings in the dictionary and connect this foreign word to its possible meanings. Instead, we develop a new approach to establish Foreign-English connections. We propose to use Inter-Lingual-Index for multilingual connections. With the help of this index, WordNets are easily and effectively connected to each other by linking the words in one WordNet to their similar meanings in the other WordNets. We use Turkish as a foreign language and generate a list of 2812 semi-automatically labeled seed words. We propose using commute time as a proximity measure with random walk model for word polarity detection. We show that besides improving the performance for Turkish, our approach also improves the performance for English.

3 Approach

3.1 Monolingual Graph Construction

We construct an undirected weighted graph G = (V, E) comprising a set V of vertices and a set E of edges. Vertices correspond to word and part-of-speech pairs in WordNet. Two words are connected with if they have one or more of the *synonym, hypernym, also see, similar to and derivation* relations in WordNet. Weight of an edge between two words is directly proportional to the number of WordNet relations between them.

3.2 Multilingual Graph Construction

Foreign WordNets are in general not as comprehensive as the English WordNet. However, most WordNets such as EuroWordNet (Vossen, 1998) and BalkaNet (Tufiş et al., 2004) are designed to be compatible with English WordNet. This compatibility provides a simple and effective way to integrate such WordNets to the powerful English WordNet. We extend our word relatedness graph by connecting the words in English WordNet with similar words in foreign WordNet by using the Inter-Lingual-Index. With the help of this index, it is possible to reach from a synset in any WordNet to the synsets of the same meaning in the other WordNets.

3.3 Random Walk with Commute Time

Consider a random walk (Lovazs, 1996) on graph G. If we are on vertex i, the probability of moving to the neighbor vertex j in the next step is directly proportional to the weight of the edge between i and j. Thus, the transition probability p_{ij} of moving from vertex i to vertex j is as follows:

$$p_{ij} = \frac{w_{ij}}{\sum_k w_{ik}}$$

Here, W_{ij} is the weight of the edge between vertices i and j, and k denotes all the neighbors of vertex i. *Hitting time* and *commute time* are two proximity measures originating from random walks. *Hitting time* between vertex i and vertex j, denoted by h_{ij}, is the expected number of steps in a random walk before vertex j is visited for the first time starting from vertex i (Sarkar , 2010). It can be calculated recursively as follows:

$$h_{ij} = \begin{cases} 0, & i = j \\ 1 + \sum_{k} p_{ik} h_{kj}, & i \neq j \end{cases}$$

where k denotes all neighbors of vertex i. Hitting time has been used to find word polarity by Hassan and Radev (2010), who have shown that it achieves the state of art performance in the literature. A drawback of hitting time is that it is not symmetric. It is possible to end up with situations where vertex i is close to vertex j (h_{ij} is small), but vertex j is far away from vertex i (h_{ji} is big). We propose using the commute time proximity measure, which is a symmetric extension of hitting time.

Commute time between vertex i and vertex j, denoted by c_{ij}, is the expected number of steps in a random walk to reach vertex j for the first time starting from vertex i and return to vertex i again. It can be calculated by using hitting time:

$$c_{ij} = h_{ij} + h_{ji}$$

Hitting and commute time are sensitive to long paths far away from the starting node (Sarkar, 2010). In general, similar words tend to be close to each other on a word relatedness graph. Therefore, we use *T-truncated hitting and commute* time, which only consider paths shorter than T.

To find the polarity of a given word, we start a random walk from that word and compute the commute time to the set of positive (P) and negative (N) seed words. Let $c_{i|P}$ be the average of truncated commute times from i to each seed in P and $c_{i|N}$ be the average of truncated commute times from i to each seed in N. If $c_{i|P}$ is less than $c_{i|N}$ word i is classified as positive, otherwise it is classified as negative. When the graph and the size of the seed list is large calculation of $c_{i|P}$ and $c_{i|N}$ is time consuming. We use a sampling approach to estimate $c_{i|P}$ and $c_{i|N}$ similar to previous works (Hassan and Radev, 2010; Sarkar, 2010).

We start M independent random walks with maximum length of T. Hitting one of the labeled seed words and returning to the starting word is the stopping condition. The length of a random walk in which the stopping condition is not met is estimated as T. Let's assume that m of M random walks met the stopping condition and the length of each random walk is $\langle t_1, t_2, \ldots, t_m \rangle$. S denotes set of positive and negative seed words. Then truncated commute time is estimated as:

$$c^{*}_{i|S} = \frac{\sum_{i=1}^{m} t_i}{M} + (1 - \frac{m}{M})T$$

The summary of our approach to find polarity of a given word is shown in Algorithm 1.

```
▪ For any given word i
▪ Start M random walks with length T on G.
▪ Calculate $c^*_{i|P}$ as estimated commute time to set of positive seeds.
▪ Start M random walks with length T on G.
▪ Calculate $c^*_{i|N}$ as estimated commute time to set of negative seeds.
▪ If $c^*_{i|P} > c^*_{i|N}$ classify word i as negative.
▪ Else classify word i as positive
```

Algorithm 1. Polarity detection using random walk model with estimated commute time

4 Experiments

We apply our approach to detect polarities of English and Turkish words. We use the WordNets of each language to construct monolingual word-relatedness graphs. A multilingual graph is obtained by connecting these graphs with the Inter-Lingual-Index. We use General Inquirer as a source for English seed words. Like in previous works (Hassan and Radev, 2010; Turney and Litman, 2003), we ignore some ambiguous words and end up with 2085 negative and 1730 positive words. Like most foreign languages, Turkish does not have a resource such as General Inquirer to obtain seed words. Algorithm 2 summarizes the semi-automated method that we propose to produce foreign seed words using the Inter-Lingual-Index. By using this algorithm, we generate 1398 positive and 1414 negative seed words for Turkish.

We use random walk model over the monolingual graphs and the English-Turkish multilingual graph to identify the polarities of words. We propose using commute

```
▪ For each word i in positive English seed words.
▪ Find all synsets in English WordNet that contain i.
▪ For each synset, find similar synset j in Foreign WordNet by using Inter-Lingual-
  Index.
▪ Select each word in synset j as a possible seed word.
▪ Repeat the same procedure for negative seeds.
▪ Process the generated foreign seed lists manually to remove the ambiguous
  words.
```

Algorithm 2. Foreign Seed Generation Algorithm

time as a proximity measure and compare it with hitting time that was shown to outperform the previous approaches for English word polarity detection by Hassan and Radev, 2010. We use 10 fold cross validation in our experiments and report the accuracies of polarity detection for the English and Turkish seed words both when the monolingual and the multilingual graphs are used.

Our experimental results are summarized in Figure 1. The proposed commute time algorithm performs similarly to the hitting time method. The accuracy for English when the monolingual graph is used is 89.7%, which is comparable to 91.1% achieved by hitting time[1]. The accuracy for Turkish when the monolingual graph is used is 86.6%, which is slightly better than 84.5% achieved by hitting time. Turkish WordNet is not as rich as English WordNet. Therefore, the accuracies for Turkish are lower than the ones for English when we use the monolingual graphs.

Figure 1 shows that the multilingual approach leads to improvements for both languages. The improvement for Turkish is more significant since we take advantage of the dense English graph. Accuracy for Turkish is improved from 86.6% to 95% with the commute time method, and it is improved from 84.5% to 95.5% with the hitting time method. Accuracy for English is improved from 89.7% to 92.3% with the commute time method, and from 91.1% to 92.8% with the hitting time method. These results demonstrate that the richness of the English WordNet is a valuable resource for Turkish word polarity detection. Interestingly, Turkish WordNet is also able to boost the performance for English word polarity detection.

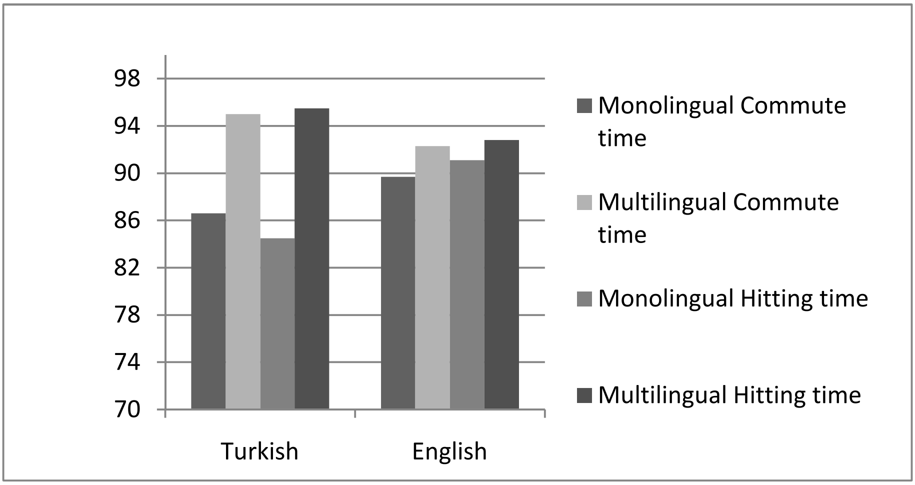

Fig. 1. Accuracies of the monolingual and multilingual approaches using commute time and hitting time methods for Turkish and English

[1] The accuracy for English when hitting time is used is reported as 93.1% in (Hassan and Radev, 2010). The difference might be due to a different version of WordNet or the seed list.

5 Conclusions

We addressed the problem of identifying the polarities of English and foreign words. Most previous studies on polarity detection focus on English and depend on language specific resources such as WordNet. Many foreign languages have WordNets. However, they are not as comprehensive as the English WordNet. In this study, we develop an approach that utilizes the compatibility of English and foreign WordNets to build a multilingual word relatedness graph. We propose using random walk model with commute time proximity measure over this graph to predict word polarities. We evaluate our approach for English and Turkish. We show that the random walk model with commute time achieves similar performance to the state of art method for English in the literature. Our multilingual approach based on connecting the English and Turkish word relatedness graphs led to significant improvement in performance for both languages.

Acknowledgements. We would like to thank Amjad Abu-Jbara and Ahmed Hassan from the University of Michigan for their assistance in providing the implementation details of their algorithm.

References

1. Morinaga, S., Yamanishi, K., Tateishi, K., Fukushima, T.: Mining Product Reputations on the Web. In: Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 341–349 (2002)
2. Turney, P.D.: Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, pp. 417–424 (2002)
3. Popescu, A., Etzioni, O.: Extracting Product Features and Opinions from Reviews. In: Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing Association for Computational Linguistics, pp. 339–346 (2005)
4. Hassan, A., Qazvinian, V., Radev, D.: What's with the Attitude? Identifying Sentences with Attitude in Online Discussions. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pp. 1245–1255 (2010)
5. Miller, G.A.: WordNet: A Lexical Database for English. Communications of the ACM 38(11), 39–41 (1995)
6. Stone, P., Dunphy, D., Smith, M., Ogilvie, D.: The General Inquirer: A Computer Approach to Content Analysis. The MIT Press, Cambridge (1966)
7. Takamura, H., Inui, T., Okumura, M.: Extracting Semantic Orientations of Words Using Spin Model. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, pp. 133–140 (2005)
8. Hassan, A., Radev, D.: Identifying Text Polarity Using Random Walks. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pp. 395–403 (2010)

9. Vossen, P.: Eurowordnet: A Multilingual Database with Lexical Semantic Networks. Kluwer Academic Publishers, Norwell (1998)
10. Tufiş, D., Cristea, D., Stamou, S.: Balkanet: Aims, Methods, Results and Perspectives. A General Overview. *Romanian* Journal on Science and Technology of Information 7, 9–43 (2004)
11. Bilgin, O., Çetinoglu, Ö., Oflazer, K.: Building a Wordnet for Turkish. Romanian Journal on Information Science and Technology 7, 163–172 (2004)
12. Turney, P.D., Littman, M.L.: Measuring Praise and Criticism: Inference of Semantic Orientation from Association. ACM Transactions on Information Systems 21(4), 315–346 (2003)
13. Hassan, A., Abu-Jbara, A., Jha, R., Radev, D.: Identifying the Semantic Orientation of Foreign Words. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, vol. 2, pp. 592–597 (2011)
14. Lovasz, L.: Random Walks on Graphs: A Survey. Bolyai Society Mathematical Studies 2, 353–398 (1996)
15. Sarkar, P.: Tractable Algorithms for Proximity Search on Large Graphs. Ph.D. Thesis, Carnegie Mellon University (2010)