Identifying the Semantic Orientation of Foreign Words

Ahmed Hassan
EECS Department
University of Michigan
Ann Arbor, MI
hassanam@umich.edu

Amjad Abu-Jbara
EECS Department
University of Michigan
Ann Arbor, MI
amjbara@umich.edu

Rahul Jha
EECS Department
University of Michigan
Ann Arbor, MI
rahuljha@umich.edu

Dragomir Radev
EECS Department and School of Information
University of Michigan
Ann Arbor, MI
radev@umich.edu

Abstract

We present a method for identifying the positive or negative semantic orientation of foreign words. Identifying the semantic orientation of words has numerous applications in the areas of text classification, analysis of product review, analysis of responses to surveys, and mining online discussions. Identifying the semantic orientation of English words has been extensively studied in literature. Most of this work assumes the existence of resources (e.g. Wordnet, seeds, etc) that do not exist in foreign languages. In this work, we describe a method based on constructing a multilingual network connecting English and foreign words. We use this network to identify the semantic orientation of foreign words based on connection between words in the same language as well as multilingual connections. The method is experimentally tested using a manually labeled set of positive and negative words and has shown very promising results.

1 Introduction

A great body of research work has focused on identifying the semantic orientation of words. Word polarity is a very important feature that has been used in several applications. For example, the problem of mining product reputation from Web reviews has been extensively studied (Turney, 2002; Morinaga et al., 2002; Nasukawa and Yi, 2003; Popescu and Etzioni, 2005; Banea et al., 2008). This is a very important task given the huge amount of product reviews written on the Web and the difficulty of manually handling them. Another interesting application is mining attitude in discussions (Hassan et al., 2010), where the attitude of participants in a discussion is inferred using the text they exchange.

Due to its importance, several researchers have addressed the problem of identifying the semantic orientation of individual words. This work has almost exclusively focused on English. Most of this work used several language dependent resources. For example Turney and Littman (2003) use the entire English Web corpus by submitting queries consisting of the given word and a set of seeds to a search engine. In addition, several other methods have used Wordnet (Miller, 1995) for connecting semantically related words (Kamps et al., 2004; Takamura et al., 2005; Hassan and Radev, 2010).

When we try to apply those methods to other languages, we run into the problem of the lack of resources in other languages when compared to English. For example, the General Inquirer lexicon (Stone et al., 1966) has thousands of English words labeled with semantic orientation. Most of the literature has used it as a source of labeled seeds or for evaluation. Such lexicons are not readily available in other languages. Another source that has been widely used for this task is Wordnet (Miller, 1995). Even though other Wordnets have been built for other languages, their coverage is very limited when compared to the English Wordnet.

In this work, we present a method for predicting the semantic orientation of foreign words. The pro-

Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics:shortpapers, pages 592–597,
Portland, Oregon, June 19-24, 2011.

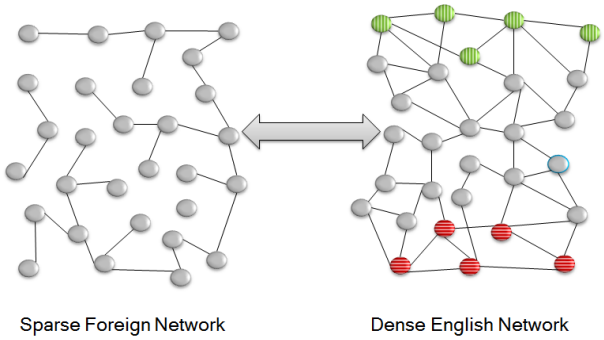

Figure 1: Sparse Foreign Networks are connected to Dense English Networks. Dashed nodes represent labeled positive and negative seeds.

posed method is based on creating a multilingual network of words that represents both English and foreign words. The network has English-English connections, as well as foreign-foreign connections and English-foreign connections. This allows us to benefit from the richness of the resources built for the English language and in the meantime utilize resources specific to foreign languages. Figure 1 shows a multilingual network where a sparse foreign network and a dense English network are connected. We then define a random walk model over the multilingual network and predict the semantic orientation of any given word by comparing the mean hitting time of a random walk starting from it to a positive and a negative set of seed English words.

We use both Arabic and Hindi for experiments. We compare the performance of several methods using the foreign language resources only and the multilingual network that has both English and foreign words. We show that bootstrapping from languages with dense resources such as English is useful for improving the performance on other languages with limited resources.

The rest of the paper is structured as follows. In section 2, we review some of the related prior work. We define our problem and explain our approach in Section 3. Results and discussion are presented in Section 4. We conclude in Section 5.

2 Related Work

The problem of identifying the polarity of individual words is a well-studied problem that attracted several research efforts in the past few years. In this section, we survey several methods that addressed this problem.

The work of Hatzivassiloglou and McKeown (1997) is among the earliest efforts that addressed this problem. They proposed a method for identifying the polarity of adjectives. Their method is based on extracting all conjunctions of adjectives from a given corpus and then they classify each conjunctive expression as either the same orientation such as "simple and well-received" or different orientation such as "simplistic but well-received". Words are clustered into two sets and the cluster with the higher average word frequency is classified as positive.

Turney and Littman (2003) identify word polarity by looking at its statistical association with a set of positive/negative seed words. They use two statistical measures for estimating association: Pointwise Mutual Information (PMI) and Latent Semantic Analysis (LSA). Co-occurrence statistics are collected by submitting queries to a search engine. The number of hits for positive seeds, negative seeds, positives seeds near the given word, and negative seeds near the given word are used to estimate the association of the given word to the positive/negative seeds.

Wordnet (Miller, 1995), thesaurus and co-occurrence statistics have been widely used to measure word relatedness by several semantic orientation prediction methods. Kamps et al. (2004) use the length of the shortest-path in Wordnet connecting any given word to positive/negative seeds to identify word polarity. Hu and Liu (2004) use Wordnet synonyms and antonyms to bootstrap from words with known polarity to words with unknown polarity. They assign any given word the label of its synonyms or the opposite label of its antonyms if any of them are known.

Kanayama and Nasukawa (2006) used syntactic features and context coherency, defined as the tendency for same polarities to appear successively, to acquire polar atoms. Takamura et al. (2005) proposed using spin models for extracting semantic orientation of words. They construct a network of words using gloss definitions, thesaurus and co-occurrence statistics. They regard each word as an electron. Each electron has a spin and each spin has a direction taking one of two values: up or down.

Two neighboring spins tend to have the same orientation from an energetic point of view. Their hypothesis is that as neighboring electrons tend to have the same spin direction, neighboring words tend to have similar polarity. Hassan and Radev (2010) use a random walk model defined over a word relatedness graph to classify words as either positive or negative. Words are connected based on Wordnet relations as well as co-occurrence statistics. They measure the random walk mean hitting time of the given word to the positive set and the negative set. They show that their method outperforms other related methods and that it is more immune to noisy word connections.

Identifying the semantic orientation of individual words is closely related to subjectivity analysis. Subjectivity analysis focused on identifying text that presents opinion as opposed to objective text that presents factual information (Wiebe, 2000). Some approaches to subjectivity analysis disregard the context phrases and words appear in (Wiebe, 2000; Hatzivassiloglou and Wiebe, 2000; Banea et al., 2008), while others take it into consideration (Riloff and Wiebe, 2003; Yu and Hatzivassiloglou, 2003; Nasukawa and Yi, 2003; Popescu and Etzioni, 2005).

3 Approach

The general goal of this work is to mine the semantic orientation of foreign words. We do this by creating a multilingual network of words. In this network two words are connected if we believe that they are semantically related. The network has English-English, English-Foreign and Foreign-Foreign connections. Some of the English words will be used as seeds for which we know the semantic orientation.

Given such a network, we will measure the mean hitting time in a random walk starting at any given word to the positive set of seeds and the negative set of seeds. Positive words will be more likely to hit the positive set faster than hitting the negative set and vice versa. In the rest of this section, we define how the multilingual word network is built and describe an algorithm for predicting the semantic orientation of any given word.

3.1 Multilingual Word Network

We build a network $G(V, E)$ where $V = V_{en} \cup V_{fr}$ is the union of a set of English and foreign words. E is a set of edges connecting nodes in V. There are three types of connections: English-English connections, Foreign-Foreign connections and English-Foreign connections.

For the English-English connections, we use Wordnet (Miller, 1995). Wordnet is a large lexical database of English. Words are grouped in synsets to express distinct concepts. We add a link between two words if they occur in the same Wordnet synset. We also add a link between two words if they have a hypernym or a similar-to relation.

Foreign-Foreign connections are created in a similar way to the English connections. Some other languages have lexical resources based on the design of the Princeton English Wordnet. For example: Euro Wordnet (EWN) (Vossen, 1997), Arabic Wordnet (AWN) (Elkateb, 2006; Black and Fellbaum, 2006; Elkateb and Fellbaum, 2006) and the Hindi Wordnet (Narayan et al., 2002; S. Jha, 2001). We also use co-occurrence statistics similar to the work of Hatzivassiloglou and McKeown (1997).

Finally, to connect foreign words to English words, we use a foreign to English dictionary. For every word in a list of foreign words, we look up its meaning in a dictionary and add an edge between the foreign word and every other English word that appeared as a possible meaning for it.

3.2 Semantic Orientation Prediction

We use the multilingual network we described above to predict the semantic orientation of words based on the mean hitting time to two sets of positive and negative seeds. Given the graph $G(V, E)$, we described in the previous section, we define the transition probability from node i to node j by normalizing the weights of the edges out from i:

$$P(j|i) = Wij / \sum_k W_{ik} \quad (1)$$

The mean hitting time $h(i|j)$ is the average number of steps a random walker, starting at i, will take to enter state j for the first time (Norris, 1997). Let the average number of steps that a random walker starting at some node i will need to enter a state

$k \in S$ be $h(i|S)$. It can be formally defined as:

$$h(i|S) = \begin{cases} 0 & i \in S \\ \sum_{j \in V} p_{ij} \times h(j|S) + 1 & \text{otherwise} \end{cases} \quad (2)$$

where p_{ij} is the transition probability between node i and node j.

Given two lists of seed English words with known polarity, we define two sets of nodes $S+$ and $S-$ representing those seeds. For any given word w, we calculate the mean hitting time between w and the two seed sets $h(w|S+)$ and $h(w|S-)$. If $h(w|S+)$ is greater than $h(w|S-)$, the word is classified as negative, otherwise it is classified as positive. We used the list of labeled seeds from (Hatzivassiloglou and McKeown, 1997) and (Stone et al., 1966). Several other similarity measures may be used to predict whether a given word is closer to the positive seeds list or the negative seeds list (e.g. average shortest path length (Kamps et al., 2004)). However hitting time has been shown to be more efficient and more accurate (Hassan and Radev, 2010) because it measures connectivity rather than distance. For example, the length of the shortest path between the words "good" and "bad" is only 5 (Kamps et al., 2004).

4 Experiments

4.1 Data

We used Wordnet (Miller, 1995) as a source of synonyms and hypernyms for linking English words in the word relatedness graph. We used two foreign languages for our experiments Arabic and Hindi. Both languages have a Wordnet that was constructed based on the design the Princeton English Wordnet. Arabic Wordnet (AWN) (Elkateb, 2006; Black and Fellbaum, 2006; Elkateb and Fellbaum, 2006) has 17561 unique words and 7822 synsets. The Hindi Wordnet (Narayan et al., 2002; S. Jha, 2001) has 56,928 unique words and 26,208 synsets.

In addition, we used three lexicons with words labeled as either positive or negative. For English, we used the General Inquirer lexicon (Stone et al., 1966) as a source of seed labeled words. The lexicon contains 4206 words, 1915 of which are positive and 2291 are negative. For Arabic and Hindi we constructed a labeled set of 300 words for each language for use in evaluation. Those sets were labeled by two native speakers of each language. We also used an Arabic-English and a Hindi-English dictionaries to generate Foreign-English links.

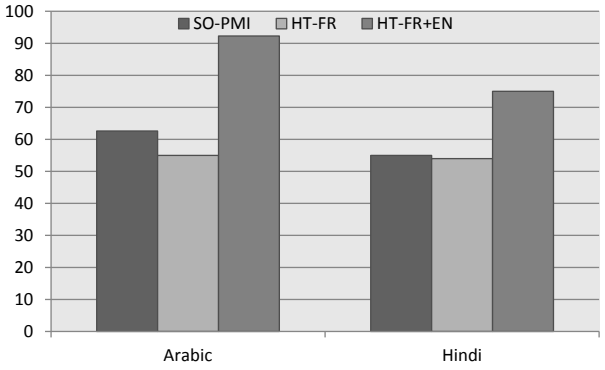

Figure 2: Accuracy of the proposed method and baselines for both Arabic and Hindi.

4.2 Results and Discussion

We performed experiments on the data described in the previous section. We compare our results to two baselines. The first is the SO-PMI method described in (Turney and Littman, 2003). This method is based on finding the semantic association of any given word to a set of positive and a set of negative words. It can be calculated as follows:

$$\text{SO-PMI}(w) = \log \frac{hits_{w,pos} \times hits_{neg}}{hits_{w,neg} \times hits_{pos}} \quad (3)$$

where w is a word with unknown polarity, $hits_{w,pos}$ is the number of hits returned by a commercial search engine when the search query is the given word and the disjunction of all positive seed words. $hits_{pos}$ is the number of hits when we search for the disjunction of all positive seed words. $hits_{w,neg}$ and $hits_{neg}$ are defined similarly. We used 7 positive and 7 negative seeds as described in (Turney and Littman, 2003).

The second baseline constructs a network of foreign words only as described earlier. It uses mean hitting time to find the semantic association of any given word. We used 10 fold cross validation for this experiment. We will refer to this system as HT-FR.

Finally, we build a multilingual network and use the hitting time as before to predict semantic orien-

tation. We used the English words from (Stone et al., 1966) as seeds and the labeled foreign words for evaluation. We will refer to this system as HT-FR + EN.

Figure 2 compares the accuracy of the three methods for Arabic and Hindi. We notice that the SO-PMI and the hitting time based methods perform poorly on both Arabic and Hindi. This is clearly evident when we consider that the accuracy of the two systems on English was 83% and 93% respectively (Turney and Littman, 2003; Hassan and Radev, 2010). This supports our hypothesis that state of the art methods, designed for English, perform poorly on foreign languages due to the limited amount of resources available in foreign languages compared to English. The figure also shows that the proposed method, which combines resources from both English and foreign languages, performs significantly better. Finally, we studied how much improvement is achieved by including links between foreign words from global Wordnets. We found out that it improves the performance by 2.5% and 4% for Arabic and Hindi respectively.

5 Conclusions

We addressed the problem of predicting the semantic orientation of foreign words. All previous work on this task has almost exclusively focused on English. Applying off-the-shelf methods developed for English to other languages does not work well because of the limited amount of resources available in foreign languages compared to English. We proposed a method based on the construction of a multilingual network that uses both language specific resources as well as the rich semantic relations available in English. We then use a model that computes the mean hitting time to a set of positive and negative seed words to predict whether a given word has a positive or a negative semantic orientation. We showed that the proposed method can predict semantic orientation with high accuracy. We also showed that it outperforms state of the art methods limited to using language specific resources.

Acknowledgments

This research was funded in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), through the U.S. Army Research Lab. All statements of fact, opinion or conclusions contained herein are those of the authors and should not be construed as representing the ofcial views or policies of IARPA, the ODNI or the U.S. Government.

References

Carmen Banea, Rada Mihalcea, and Janyce Wiebe. 2008. A bootstrapping method for building subjectivity lexicons for languages with scarce resources. In *LREC'08*.

Elkateb S. Rodriguez H Alkhalifa M. Vossen P. Pease A. Black, W. and C. Fellbaum. 2006. Introducing the arabic wordnet project. In *Third International WordNet Conference*.

Black. W. Rodriguez H Alkhalifa M. Vossen P. Pease A. Elkateb, S. and C. Fellbaum. 2006. Building a wordnet for arabic. In *Fifth International Conference on Language Resources and Evaluation*.

Black W. Vossen P. Farwell D. Rodrguez H. Pease A. Alkhalifa M. Elkateb, S. 2006. Arabic wordnet and the challenges of arabic. In *Arabic NLP/MT Conference*.

Ahmed Hassan and Dragomir Radev. 2010. Identifying text polarity using random walks. In *ACL'10*.

Ahmed Hassan, Vahed Qazvinian, and Dragomir Radev. 2010. 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*, pages 1245–1255.

Vasileios Hatzivassiloglou and Kathleen R. McKeown. 1997. Predicting the semantic orientation of adjectives. In *EACL'97*, pages 174–181.

Vasileios Hatzivassiloglou and Janyce Wiebe. 2000. Effects of adjective orientation and gradability on sentence subjectivity. In *COLING*, pages 299–305.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *KDD'04*, pages 168–177.

Jaap Kamps, Maarten Marx, Robert J. Mokken, and Maarten De Rijke. 2004. Using wordnet to measure semantic orientations of adjectives. In *National Institute for*, pages 1115–1118.

Hiroshi Kanayama and Tetsuya Nasukawa. 2006. Fully automatic lexicon expansion for domain-oriented sentiment analysis. In *EMNLP'06*, pages 355–363.

George A. Miller. 1995. Wordnet: a lexical database for english. *Commun. ACM*, 38(11):39–41.

Satoshi Morinaga, Kenji Yamanishi, Kenji Tateishi, and Toshikazu Fukushima. 2002. Mining product reputations on the web. In *KDD'02*, pages 341–349.

Dipak Narayan, Debasri Chakrabarti, Prabhakar Pande, and P. Bhattacharyya. 2002. An experience in building the indo wordnet - a wordnet for hindi. In *First International Conference on Global WordNet*.

Tetsuya Nasukawa and Jeonghee Yi. 2003. Sentiment analysis: capturing favorability using natural language processing. In *K-CAP '03: Proceedings of the 2nd international conference on Knowledge capture*, pages 70–77.

J. Norris. 1997. Markov chains. Cambridge University Press.

Ana-Maria Popescu and Oren Etzioni. 2005. Extracting product features and opinions from reviews. In *HLT-EMNLP'05*, pages 339–346.

Ellen Riloff and Janyce Wiebe. 2003. Learning extraction patterns for subjective expressions. In *EMNLP'03*, pages 105–112.

P. Pande P. Bhattacharyya S. Jha, D. Narayan. 2001. A wordnet for hindi. In *International Workshop on Lexical Resources in Natural Language Processing*.

Philip Stone, Dexter Dunphy, Marchall Smith, and Daniel Ogilvie. 1966. The general inquirer: A computer approach to content analysis. *The MIT Press*.

Hiroya Takamura, Takashi Inui, and Manabu Okumura. 2005. Extracting semantic orientations of words using spin model. In *ACL'05*, pages 133–140.

Peter Turney and Michael Littman. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems*, 21:315–346.

Peter D. Turney. 2002. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *ACL'02*, pages 417–424.

P. Vossen. 1997. Eurowordnet: a multilingual database for information retrieval. In *DELOS workshop on Cross-language Information Retrieval*.

Janyce Wiebe. 2000. Learning subjective adjectives from corpora. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence*, pages 735–740.

Hong Yu and Vasileios Hatzivassiloglou. 2003. Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences. In *EMNLP'03*, pages 129–136.