Bootstrapping Sentiment Labels For Unannotated Documents With Polarity PageRank

Christian Scheible, Hinrich Schütze

Institute for Natural Language Processing
Universität Stuttgart
Germany
scheibcn@ims.uni-stuttgart.de

Abstract

We present a novel graph-theoretic method for the initial annotation of high-confidence training data for bootstrapping sentiment classifiers. We estimate polarity using *topic-specific PageRank*. Sentiment information is propagated from an initial seed lexicon through a joint graph representation of words and documents. We report improved classification accuracies across multiple domains for the base models and the maximum entropy model bootstrapped from the PageRank annotation.

Keywords: sentiment analysis, self-training, graphs

1. Introduction

Recognizing the sentiment expressed in a document is an important task in natural language processing. Semi- and unsupervised methods are a promising approach to this task because they reduce the need for expensive manual labeling. In this paper, we introduce a new method for bootstrapping a sentiment classifier from a seed lexicon. We first apply *topic-specific PageRank* to a graph of both words and documents. The resulting polarity annotation of the documents is then used to train a high-performance maximum entropy classifier on the documents. Methods based on lexical resources are popular in Sentiment analysis (e.g. Taboada et al. (2011)). This serves as motivation for our method: words on the one hand are sufficient for decent (though not optimal) sentiment classification performance; on the other hand, words are a good feature space for semisupervised sentiment classification because many words are strong features and the size of the feature space is manageable (compared to, say, bigrams). A high-performance statistical classifier can then be trained on the larger space of all available features, resulting in higher performance than the initial PageRank classifier.

This paper makes the following contributions: (i) We introduce Polarity PageRank (PPR), a new semi-supervised sentiment classifier that integrates lexicon induction with document classification. (ii) We show that PPR can be successfully applied to sentiment classification by evaluating it on an English reference corpus. (iii) We show that classification accuracy on documents can be further improved by training a more sophisticated classifier that takes advantage of all available features on the automatically labeled documents.

In the following sections, we discuss related work, introduce graph structure and algorithm and evaluate the performance of our method.

2. Related Work

Turney (2002) induces a polarity lexicon by measuring the association of terms with a set of seed words whose polarity is known. The resulting lexicon is used for classifying reviews by calculating the average polarity of each document. Turney concludes that these averages are highly correlated to the actual polarity of the documents.

Wiebe and Riloff (2005) introduce a bootstrapping approach for subjectivity classification that learns patterns of subjectivity clues from unannotated texts. These clues serve as a source for a Naive Bayes classifier that produces additional high-confidence input for the pattern learner. Initial rules need to be hand-crafted which requires linguistic expert knowledge about a language.

He (2010) presents a self-training approach for review classification. The importance of each lexicon item is taken into account and is estimated from the unlabeled texts. This leads to an increase of accuracy in review classification. The focus of He's work is on correctly estimating the importance of each feature for sentiment classification.

Hassan and Radev (2010) induce a polarity lexicon by constructing a graph that links words using WordNet relations like hypernymy. On these graphs polarity is propagated using a random walk model that handles positive and negative words separately. Their approach outperforms a state of the art method on the task of assessing the polarity of features. In contrast, we adopt document classification accuracy as our evaluation metric.

We know of no other work that formalizes sentiment classification in the framework of a joint graph of words and documents. A second innovation that distinguishes our approach is that the induced lexicon is extended to the entire feature space using a sophisticated classifier.

3. Polarity PageRank

3.1. Word Graphs

Hatzivassiloglou and McKeown (1997) introduced a novel way of calculating the polarities of words. In their approach, words are represented as nodes in a graph. Links between the nodes denote some type of relationship between the words. An example for such a graph is provided in Figure 1. In the original paper, words coordinated with

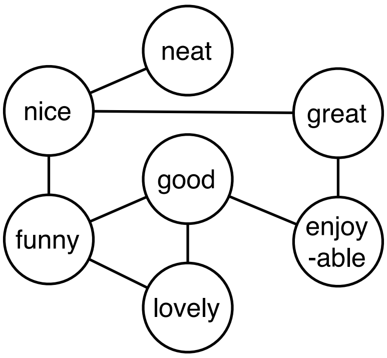

Figure 1: Word graph

and are assumed to share polarities while words coordinated with *but* are expected to have opposing polarities. We adopt this approach for the word-word links in our graph, but only use *and*-links.

3.2. The Polarity PageRank Method

Polarity PageRank (PPR) is derived from topic-specific PageRank (Haveliwala, 2003), a method that calculates the link-based "authority" of a page on the world wide web in relation to a specific topic rather than its overall importance. The method differs from standard PageRank by increasing the teleporting probability for pages in the right topic, thereby boosting the importance of links that emanate from those pages.

The PageRank of a node (a page in web information retrieval, a word or document in our case) is its value in the dominant left eigenvector of the transition probability matrix M of the underlying graph (the link graph of all pages in web information retrieval, the sentiment graph linking words and documents in our case). We refer to the dominant left eigenvector as the rank vector ($\vec{r}$). It can be computed as the fixed point of the following equation, e.g., by the power method:

$$\vec{r} = \vec{r} \times M \tag{1}$$

Teleportation can be modeled in the graph M directly by incorporating teleportation into the edge weights; or indirectly through a "raw" probability transition matrix A without teleportation probabilities and a teleportation vector $\vec{t}$, which can then be used together to calculate M (Haveliwala, 2003).

In standard PageRank (Page et al., 1998), $\vec{t}$ is the uniform distribution: all nodes are equally likely to be teleportation targets. Topic-specific PageRank gives a higher weight in $\vec{t}$ to nodes that are good representatives of a certain topic, making these nodes more likely to be teleportation targets.

The idea behind PPR is to view positive and negative as two different topics in topic-specific PageRank. Thus, PPR combines two independent runs (*positive* and *negative*) of topic-specific PageRank. In the *positive* (resp., *negative*) run, $\vec{t_p}$ (resp., $\vec{t_n}$) is defined so as to give high weights to nodes whose polarity is known to be *positive* (resp., *negative*). The entry t_{xi} for word w_i of the teleportation vector t_x for the class x is defined as follows:

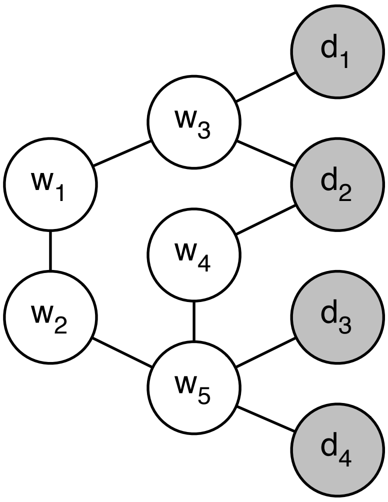

Figure 2: Word-document graph

$$t_{xi} = \begin{cases} \dfrac{1}{n_x} & \text{if } w_i \text{ in class x} \\ 0 & \text{if } w_i \text{ not in class x} \end{cases}$$

where n_x is the number of words in the class x in the polarity lexicon and n the number of words in the graph. $x \in \{\text{positive, negative}\}$.

The initial assignments to positive or negative polarity are taken from an existing polarity lexicon. There are, however, no limitations concerning the size of the lexicon; this makes it possible to start with a small set of initial polarities. From $\vec{t_p}$ (resp., $\vec{t_n}$), we construct the matrix M_p (resp., M_n). With M_p (resp., M_n), we calculate the positive rank vector $\vec{r_p}$ (resp., the negative rank vector $\vec{r_n}$) applying Equation 1. These vectors will contain high scores for words that are important to the respective classes. The sets of positive (*pos*) and negative (*neg*) words are defined as follows:

$$pos = \{v : r_p(v) \geq r_n(v)\} \tag{2}$$

$$neg = \{v : r_n(v) > r_p(v)\} \tag{3}$$

In the resulting lexicon, each word in the graph is either positive or negative.

PPR bears some resemblance to the method of Hassan and Radev (2010) discussed earlier. However, it is simpler, using standard eigenvector computations that are available in any numerical software library, and it is also more efficient, avoiding the need for expensive Monte Carlo sampling. Computational simplicity and efficiency are of particular importance for the much larger graph we are working with – a graph that contains both words and documents.

3.3. Document Classification with Polarity PageRank

Common methods for applying a polarity lexicon to document classification are the calculation of a lexical score that counts the number of positive and negative words in some way (e.g. Turney (2002)); using lexical scores as hard-coded features for a more sophisticated classifier (e.g. Melville et al. (2009)); and a mixture of both approaches (e.g. He (2010)).

Method	Books	DVD	Electronics	Kitchen	Overall
AP (base)	51.2	58.5	57.1	56.6	55.8
PPR (base)	**68.7**	**67.7**	**67.1**	**67.6**	**67.8**
AP (MaxEnt)	70.4	69.6	74.1	75.7	72.8
PPR (MaxEnt)	**71.1**	**72.5**	**76.7**	**80.3**	**75.6**

Table 1: Classifier accuracies

PPR, the novel approach we introduce in this paper, integrates lexicon induction and lexicon application in one unified formalism. This way of formalizing the problem is not unlike many information retrieval methods that also view words and documents as the same formal object (e.g., (Turtle and Croft, 1991)). This way, the relations between documents and the words occurring in them is modeled reflects structures commonly used in information retrieval.

To realize the joint graph structure, the concept of word graphs (see Section 3.1.) needs to be extended to include documents. We will refer to these graphs as *word-document graphs*. Each document in the word-document graph is linked to each word that occurs in it. The degree of association between a word w and a document d is given by their normalized term frequency tf (Salton and McGill, 1983):

$$\text{tf}_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

where $n_{i,j}$ is the occurrence count of term t_i in document d_j and $|D|$ the size of the collection, and $|\{d : t_i \in d\}|$ the number of documents that contain the term t_i.

In contrast to the standard setup of PageRank in document retrieval, the documents are not linked directly to each other. Instead, document nodes are linked only to word nodes. However, word nodes can be linked to each other. This way, the relationships between documents are defined through the relationships of their terms. The relations necessary for these links do not have to be obtained from the documents in the graph but can be gathered from external sources as well (as we do). Figure 2 contains an example graph with words $w_1 \dots w_5$ and documents $d_1 \dots d_4$. By including document nodes in the graph, the documents themselves can be labeled with polarity using PPR. We simply use Equation 2 for the joint graph, setting the teleportation probabilities for documents to 0. Confidence is assessed by the log ratio of r_p and r_n.

In addition, these graphs can easily be made bilingual. If a word-document graph exists in two languages A and B, the graphs can be combined by adding links from a standard bilingual dictionary that translates between the languages (cf. Scheible et al. (2010)).

4. Bootstrapping

Self-training is a machine-learning technique for increasing the amount of training data for a classifier through annotating unlabeled data. In a typical setting, a small set of labeled data and a large set of unlabeled data are available. A classifier is then trained on the labeled data and applied to the unlabeled data.

One version of self-training is to provide an initial classifier instead of an initial labeled dataset. This version has the advantage that it can be applied to a corpus that is completely unlabeled.

The motivation behind this model is that we can produce a base classifier and label a small set of documents with the highest confidence on the word level using a dictionary. These documents and their annotations are training input for a more sophisticated classifier that uses additionally available features from a higher linguistic level in the dataset.

The approach in this paper combines initial document classification using PPR from which a high-confidence partition is used for subsequent bootstrapping steps with a maximum entropy classifier.

5. Evaluation

5.1. Baseline

The baseline against which we evaluate PPR is average polarity (Turney, 2002). The average polarity pol of all words w in a document d is

$$\text{pol}(d) = \sum_{w \in d} \frac{\text{pol}_l(w)}{|\{w : \text{pol}_l(w) \neq 0\}|},$$

where pol_l is the lexical polarity taken from a polarity lexicon. The classification confidence c of a document d is assessed through $c(d) = |pol(d)|$.

5.2. Experiments

Experiments are carried out on the Multi-Domain Sentiment Dataset (Blitzer et al., 2007). For each category, all available reviews (positive, negative, and unlabeled) are merged into a joint collection. Since we do not use the available document labels for classification, we can evaluate on the complete dataset.

We extracted adjectives coordinated with *and* from the English Wikipedia by applying simple part-of-speech search patterns. The edges for the resulting graph are weighted by coordination occurrence counts.

We use (Wilson et al., 2005)'s English polarity lexicon. We discard adjectives that were not found in Wikipedia coordinations. Thus, we only use part of the lexicon, a total of 2174 words. The high-performance classifer in our experiment is the Stanford MaxEnt classifier (Manning and Klein, 2003).

5.3. Experiments and Results

We tried two variations of Polarity PageRank. In the first version, some words of the graph are labeled and the polarities of the documents are calculated in one PageRank run.

The second version first calculates the positive and negative eigenvector on the word graph and then uses them as teleportation vectors on the word-document graph.

The top section of Table 1 shows the accuracy of the average polarity (AP) and Polarity PageRank (PPR) base classifiers before bootstrapping on the different corpora splits. The overall accuracies in the last column are averages over the four domains calculated by taking the number of documents per domain into account.

From each of the base classifiers, we select the 15% of the documents with the highest confidence as training data for individual bootstrapping instances. In each bootstrapping iteration, the top 1% with the highest maximum entropy classifier confidence is selected as additional training data. The accuracies of the resulting classifiers are listed in the bottom section of Table 1.

5.4. Discussion

Taking a look at the best-performing classifier for English, we can see that accuracy can be gained from using Polarity PageRank instead of average polarities. We achieve higher classification accuracies on all domains.

The improvements of PPR maximum entropy models over AP models is smaller than the margin of the base models which is due to the availability of more features to the classifiers.

6. Conclusion and Future Work

We have introduced Polarity PageRank, a new semi-supervised sentiment classifier that integrates lexicon induction with document classification in one unified graph-theoretic formalism. We have shown that PPR outperforms a baseline classifier and that its performance can be further improved by a bootstrapping method that can take advantage of the entire feature space available.

We were able to show that Polarity PageRank-based annotation improved results over annotation with average polarity. The high accuracy improvements translate to increased performance of subsequently trained maximum entropy classifiers across all domains. and for a joint domain model.

In future work, we will attempt to include more sophisticated features in the graph, e.g., negation-based features or bigrams. It is an open question if the PPR would be as accurate as bootstrapping if the full feature set could be made available to PPR. In addition our approach can be extended to multilingual sources by using multiple word-document graphs and a bilingual dictionary.

7. Acknowledgements

This work was funded by the DFG as part of the SFB 732 project D7.

8. References

J. Blitzer, M. Dredze, and F. Pereira. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, volume 45.

A. Hassan and D. Radev. 2010. Identifying text polarity using random walks. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 395–403. Association for Computational Linguistics.

V. Hatzivassiloglou and K.R. McKeown. 1997. Predicting the semantic orientation of adjectives. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pages 174–181. Association for Computational Linguistics.

T.H. Haveliwala. 2003. Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search. *IEEE transactions on knowledge and data engineering*, pages 784–796.

Y. He. 2010. Learning sentiment classification model from labeled features. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1685–1688. ACM.

C. Manning and D. Klein. 2003. Optimization, maxent models, and conditional estimation without magic. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: Tutorials-Volume 5*, page 8. Association for Computational Linguistics.

P. Melville, W. Gryc, and R.D. Lawrence. 2009. Sentiment analysis of blogs by combining lexical knowledge with text classification. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1275–1284. ACM.

L. Page, S. Brin, R. Motwani, and T. Winograd. 1998. The pagerank citation ranking: Bringing order to the web. Technical report, Technical report, Stanford Digital Library Technologies Project, 1998.

G. Salton and M.J. McGill. 1983. Introduction to modern information retrieval. *New York*.

Christian Scheible, Florian Laws, Lukas Michelbacher, and Hinrich Schütze. 2010. Sentiment translation through multi-edge graphs. In *Coling 2010: Posters*, pages 1104–1112, Beijing, China, August. Coling 2010 Organizing Committee.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37:267–307.

P.D. Turney. 2002. 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*, pages 417–424. Association for Computational Linguistics.

Howard Turtle and W. Bruce Croft. 1991. Evaluation of an inference network-based retrieval model. *ACM Transactions on Information Systems*, 9(3):187–222.

J. Wiebe and E. Riloff. 2005. Creating subjective and objective sentence classifiers from unannotated texts. *Computational Linguistics and Intelligent Text Processing*, pages 486–497.

T. Wilson, J. Wiebe, and P. Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 347–354. Association for Computational Linguistics.