Discovering the Discriminative Views: Measuring Term Weights for Sentiment Analysis

Jungi Kim, Jin-Ji Li and Jong-Hyeok Lee
Division of Electrical and Computer Engineering
Pohang University of Science and Technology, Pohang, Republic of Korea
{yangpa,ljj,jhlee}@postech.ac.kr

Abstract

This paper describes an approach to utilizing term weights for sentiment analysis tasks and shows how various term weighting schemes improve the performance of sentiment analysis systems. Previously, sentiment analysis was mostly studied under data-driven and lexicon-based frameworks. Such work generally exploits textual features for fact-based analysis tasks or lexical indicators from a sentiment lexicon. We propose to model term weighting into a sentiment analysis system utilizing collection statistics, contextual and topic-related characteristics as well as opinion-related properties. Experiments carried out on various datasets show that our approach effectively improves previous methods.

1 Introduction

With the explosion in the amount of commentaries on current issues and personal views expressed in weblogs on the Internet, the field of studying how to analyze such remarks and sentiments has been increasing as well. The field of opinion mining and sentiment analysis involves extracting opinionated pieces of text, determining the polarities and strengths, and extracting holders and targets of the opinions.

Much research has focused on creating testbeds for sentiment analysis tasks. Most notable and widely used are Multi-Perspective Question Answering (MPQA) and Movie-review datasets. MPQA is a collection of newspaper articles annotated with opinions and private states at the subsentence level (Wiebe et al., 2003). Movie-review dataset consists of positive and negative reviews from the Internet Movie Database (IMDb) archive (Pang et al., 2002).

Evaluation workshops such as TREC and NTCIR have recently joined in this new trend of research and organized a number of successful meetings. At the TREC Blog Track meetings, researchers have dealt with the problem of retrieving topically-relevant blog posts and identifying documents with opinionated contents (Ounis et al., 2008). NTCIR Multilingual Opinion Analysis Task (MOAT) shared a similar mission, where participants are provided with a number of topics and a set of relevant newspaper articles for each topic, and asked to extract opinion-related properties from enclosed sentences (Seki et al., 2008).

Previous studies for sentiment analysis belong to either the data-driven approach where an annotated corpus is used to train a machine learning (ML) classifier, or to the lexicon-based approach where a pre-compiled list of sentiment terms is utilized to build a sentiment score function.

This paper introduces an approach to the sentiment analysis tasks with an emphasis on how to represent and evaluate the weights of sentiment terms. We propose a number of characteristics of good sentiment terms from the perspectives of informativeness, prominence, topic–relevance, and semantic aspects using collection statistics, contextual information, semantic associations as well as opinion–related properties of terms. These term weighting features constitute the sentiment analysis model in our opinion retrieval system. We test our opinion retrieval system with TREC and NTCIR datasets to validate the effectiveness of our term weighting features. We also verify the effectiveness of the statistical features used in data-driven approaches by evaluating an ML classifier with labeled corpora.

2 Related Work

Representing text with salient features is an important part of a text processing task, and there exists many works that explore various features for

Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, pages 253–261,
Suntec, Singapore, 2-7 August 2009.

text analysis systems (Sebastiani, 2002; Forman, 2003). Sentiment analysis task have also been using various lexical, syntactic, and statistical features (Pang and Lee, 2008). Pang et al. (2002) employed n-gram and POS features for ML methods to classify movie-review data. Also, syntactic features such as the dependency relationship of words and subtrees have been shown to effectively improve the performances of sentiment analysis (Kudo and Matsumoto, 2004; Gamon, 2004; Matsumoto et al., 2005; Ng et al., 2006).

While these features are usually employed by data-driven approaches, there are unsupervised approaches for sentiment analysis that make use of a set of terms that are semantically oriented toward expressing subjective statements (Yu and Hatzivassiloglou, 2003). Accordingly, much research has focused on recognizing terms' semantic orientations and strength, and compiling sentiment lexicons (Hatzivassiloglou and Mckeown, 1997; Turney and Littman, 2003; Kamps et al., 2004; Whitelaw et al., 2005; Esuli and Sebastiani, 2006).

Interestingly, there are conflicting conclusions about the usefulness of the statistical features in sentiment analysis tasks (Pang and Lee, 2008). Pang et al. (2002) presents empirical results indicating that using term presence over term frequency is more effective in a data-driven sentiment classification task. Such a finding suggests that sentiment analysis may exploit different types of characteristics from the topical tasks, that, unlike fact-based text analysis tasks, repetition of terms does not imply a significance on the overall sentiment. On the other hand, Wiebe et al. (2004) have noted that hapax legomena (terms that only appear once in a collection of texts) are good signs for detecting subjectivity. Other works have also exploited rarely occurring terms for sentiment analysis tasks (Dave et al., 2003; Yang et al., 2006).

The opinion retrieval task is a relatively recent issue that draws both the attention of IR and NLP communities. Its task is to find relevant documents that also contain sentiments about a given topic. Generally, the opinion retrieval task has been approached as a two–stage task: first, retrieving topically relevant documents, then reranking the documents by the opinion scores (Ounis et al., 2006). This approach is also appropriate for evaluation systems such as NTCIR MOAT that assumes that the set of topically relevant documents are already known in advance. On the other hand, there are also some interesting works on modeling the topic and sentiment of documents in a unified way (Mei et al., 2007; Zhang and Ye, 2008).

3 Term Weighting and Sentiment Analysis

In this section, we describe the characteristics of terms that are useful in sentiment analysis, and present our sentiment analysis model as part of an opinion retrieval system and an ML sentiment classifier.

3.1 Characteristics of Good Sentiment Terms

This section examines the qualities of useful terms for sentiment analysis tasks and corresponding features. For the sake of organization, we categorize the sources of features into either global or local knowledge, and either topic-independent or topic-dependent knowledge.

Topic-independently speaking, a good sentiment term is *discriminative* and *prominent*, such that the appearance of the term imposes greater influence on the judgment of the analysis system. The rare occurrence of terms in document collections has been regarded as a very important feature in IR methods, and effective IR models of today, either explicitly or implicitly, accommodate this feature as an Inverse Document Frequency (IDF) heuristic (Fang et al., 2004). Similarly, prominence of a term is recognized by the frequency of the term in its local context, formulated as Term Frequency (TF) in IR.

If a topic of the text is known, terms that are relevant and descriptive of the subject should be regarded to be more useful than topically-irrelevant and extraneous terms. One way of measuring this is using associations between the query and terms. Statistical measures of associations between terms include estimations by the co-occurrence in the whole collection, such as Point-wise Mutual Information (PMI) and Latent Semantic Analysis (LSA). Another method is to use proximal information of the query and the word, using syntactic structure such as dependency relations of words that provide the graphical representation of the text (Mullen and Collier, 2004). The minimum spans of words in such graph may represent their associations in the text. Also, the distance between words in the local context or in the thesaurus-like dictionaries such as WordNet may be approximated as such measure.

3.2 Opinion Retrieval Model

The goal of an opinion retrieval system is to find a set of opinionated documents that are relevant to a given topic. We decompose the opinion retrieval system into two tasks: the topical retrieval task and the sentiment analysis task. This two-stage approach for opinion retrieval has been taken by many systems and has been shown to perform well (Ounis et al., 2006). The topic and the sentiment aspects of the opinion retrieval task are modeled separately, and linearly combined together to produce a list of topically-relevant and opinionated documents as below.

$$Score_{OpRet}(D,Q) = \lambda \cdot Score_{rel}(D,Q) + (1-\lambda) \cdot Score_{op}(D,Q)$$

The topic-relevance model $Score_{rel}$ may be substituted by any IR system that retrieves relevant documents for the query Q. For tasks such as NTCIR MOAT, relevant documents are already known in advance and it becomes unnecessary to estimate the relevance degree of the documents. We focus on modeling the sentiment aspect of the opinion retrieval task, assuming that the topic-relevance of documents is provided in some way.

To assign documents with sentiment degrees, we estimate the probability of a document D to generate a query Q and to possess opinions as indicated by a random variable Op.[1] Assuming uniform prior probabilities of documents D, query Q, and Op, and conditional independence between Q and Op, the opinion score function reduces to estimating the generative probability of Q and Op given D.

$$Score_{op}(D,Q) \equiv p(D \mid Op, Q) \propto p(Op, Q \mid D)$$

If we regard that the document D is represented as a bag of words and that the words are uniformly distributed, then

$$\begin{aligned} p(Op, Q \mid D) &= \sum_{w \in D} p(Op, Q \mid w) \cdot p(w \mid D) \\ &= \sum_{w \in D} p(Op \mid w) \cdot p(Q \mid w) \cdot p(w \mid D) \end{aligned} \quad (1)$$

Equation 1 consists of three factors: the probability of a word to be opinionated ($P(Op|w)$), the likelihood of a query given a word ($P(Q|w)$), and the probability of a document generating a word ($P(w|D)$). Intuitively speaking, the probability of a document embodying topically related opinion is estimated by accumulating the probabilities of all words from the document to have sentiment meanings and associations with the given query.

In the following sections, we assess the three factors of the sentiment models from the perspectives of term weighting.

3.2.1 Word Sentiment Model

Modeling the sentiment of a word has been a popular approach in sentiment analysis. There are many publicly available lexicon resources. The size, format, specificity, and reliability differ in all these lexicons. For example, lexicon sizes range from a few hundred to several hundred thousand. Some lexicons assign real number scores to indicate sentiment orientations and strengths (i.e. probabilities of having positive and negative sentiments) (Esuli and Sebastiani, 2006) while other lexicons assign discrete classes (weak/strong, positive/negative) (Wilson et al., 2005). There are manually compiled lexicons (Stone et al., 1966) while some are created semi-automatically by expanding a set of seed terms (Esuli and Sebastiani, 2006).

The goal of this paper is not to create or choose an appropriate sentiment lexicon, but rather it is to discover useful term features other than the sentiment properties. For this reason, one sentiment lexicon, namely SentiWordNet, is utilized throughout the whole experiment.

SentiWordNet is an automatically generated sentiment lexicon using a semi-supervised method (Esuli and Sebastiani, 2006). It consists of WordNet synsets, where each synset is assigned three probability scores that add up to 1: positive, negative, and objective.

These scores are assigned at sense level (synsets in WordNet), and we use the following equations to assess the sentiment scores at the word level.

$$\begin{aligned} p(Pos \mid w) &= \max_{s \in synset(w)} SWN_{Pos}(s) \\ p(Neg \mid w) &= \max_{s \in synset(w)} SWN_{Neg}(s) \\ p(Op \mid w) &= \max\left(p(Pos \mid w), p(Neg \mid w)\right) \end{aligned}$$

where $synset(w)$ is the set of synsets of w and $SWN_{Pos}(s)$, $SWN_{Neg}(s)$ are positive and negative scores of a synset in SentiWordNet. We assess the subjective score of a word as the maximum value of the positive and the negative scores, because a word has either a positive or a negative sentiment in a given context.

The word sentiment model can also make use of other types of sentiment lexicons. The sub-

[1] Throughout this paper, Op indicates $Op = 1$.

jectivity lexicon used in OpinionFinder[2] is compiled from several manually and automatically built resources. Each word in the lexicon is tagged with the strength (*strong*/*weak*) and polarity (*Positive*/*Negative*/*Neutral*). The word sentiment can be modeled as below.

$$P(Pos|w) = \begin{cases} 1.0 & \text{if w is } \textit{Positive and Strong} \\ 0.5 & \text{if w is } \textit{Positive and Weak} \\ 0.0 & \text{otherwise} \end{cases}$$

$$P(Op \mid w) = \max\left(p(Pos \mid w), p(Neg \mid w)\right)$$

3.2.2 Topic Association Model

If a topic is given in the sentiment analysis, terms that are closely associated with the topic should be assigned heavy weighting. For example, sentiment words such as *scary* and *funny* are more likely to be associated with topic words such as *book* and *movie* than *grocery* or *refrigerator*.

In the topic association model, $p(Q \mid w)$ is estimated from the associations between the word w and a set of query terms Q.

$$p(Q \mid w) = \frac{\sum_{q \in Q} \textit{Asc-Score}(q, w)}{\mid Q \mid} \propto \sum_{q \in Q} \textit{Asc-Score}(q, w)$$

$\textit{Asc-Score}(q, w)$ is the association score between q and w, and $\mid Q \mid$ is the number of query words.

To measure associations between words, we employ statistical approaches using document collections such as LSA and PMI, and local proximity features using the distance in dependency trees or texts.

Latent Semantic Analysis (LSA) (Landauer and Dumais, 1997) creates a semantic space from a collection of documents to measure the semantic relatedness of words. Point-wise Mutual Information (PMI) is a measure of associations used in information theory, where the association between two words is evaluated with the joint and individual distributions of the two words. PMI-IR (Turney, 2001) uses an IR system and its search operators to estimate the probabilities of two terms and their conditional probabilities. Equations for association scores using LSA and PMI are given below.

$$\textit{Asc-Score}_{LSA}(w_1, w_2) = \frac{1 + LSA(w_1, w_2)}{2}$$

$$\textit{Asc-Score}_{PMI}(w_1, w_2) = \frac{1 + \textit{PMI-IR}(w_1, w_2)}{2}$$

[2]http://www.cs.pitt.edu/mpqa/

For the experimental purpose, we used publicly available online demonstrations for LSA and PMI. For LSA, we used the online demonstration mode from the Latent Semantic Analysis page from the University of Colorado at Boulder.[3] For PMI, we used the online API provided by the CogWorks Lab at the Rensselaer Polytechnic Institute.[4]

Word associations between two terms may also be evaluated in the local context where the terms appear together. One way of measuring the proximity of terms is using the syntactic structures. Given the dependency tree of the text, we model the association between two terms as below.

$$\textit{Asc-Score}_{DTP}(w_1, w_2) = \begin{cases} 1.0 & \text{min. span in dep. tree} \leq D_{syn} \\ 0.5 & \text{otherwise} \end{cases}$$

where, D_{syn} is arbitrarily set to 3.

Another way is to use co-occurrence statistics as below.

$$\textit{Asc-Score}_{WP}(w_1, w_2) = \begin{cases} 1.0 & \text{if distance between} w_1 \text{and} w_2 \leq K \\ 0.5 & \text{otherwise} \end{cases}$$

where K is the maximum window size for the co-occurrence and is arbitrarily set to 3 in our experiments.

The statistical approaches may suffer from data sparseness problems especially for named entity terms used in the query, and the proximal clues cannot sufficiently cover all term–query associations. To avoid assigning zero probabilities, our topic association models assign 0.5 to word pairs with no association and 1.0 to words with perfect association.

Note that proximal features using co-occurrence and dependency relationships were used in previous work. For opinion retrieval tasks, Yang et al. (2006) and Zhang and Ye (2008) used the co-occurrence of a query word and a sentiment word within a certain window size. Mullen and Collier (2004) manually annotated named entities in their dataset (i.e. title of the record and name of the artist for music record reviews), and utilized presence and position features in their ML approach.

3.2.3 Word Generation Model

Our word generation model $p(w \mid d)$ evaluates the prominence and the discriminativeness of a word

[3]http://lsa.colorado.edu/, default parameter settings for the semantic space (TASA, 1st year college level) and number of factors (300).

[4]http://cwl-projects.cogsci.rpi.edu/msr/, PMI-IR with the Google Search Engine.

w in a document d. These issues correspond to the core issues of traditional IR tasks. IR models, such as Vector Space (VS), probabilistic models such as BM25, and Language Modeling (LM), albeit in different forms of approach and measure, employ heuristics and formal modeling approaches to effectively evaluate the relevance of a term to a document (Fang et al., 2004). Therefore, we estimate the word generation model with popular IR models' the relevance scores of a document d given w as a query.[5]

$$p(w \mid d) \equiv IR\text{-}SCORE(w, d)$$

In our experiments, we use the Vector Space model with Pivoted Normalization (VS), Probabilistic model (BM25), and Language modeling with Dirichlet Smoothing (LM).

$$VSPN(w,d) = \frac{1 + ln(1 + ln(c(w,d)))}{(1-s) + s \cdot \frac{\mid d \mid}{avgdl}} \cdot ln\frac{N+1}{df(w)}$$

$$BM25(w,d) = ln\frac{N - df(w) + 0.5}{df(w) + 0.5} \cdot \frac{(k_1+1) \cdot c(w,d)}{k_1\left((1-b) + b\frac{|d|}{avgdl}\right) + c(w,d)}$$

$$LMDI(w,d) = ln\left(1 + \frac{c(w,d)}{\mu \cdot c(w,C)}\right) + ln\frac{\mu}{\mid d \mid + \mu}$$

$c(w,d)$ is the frequency of w in d, $\mid d \mid$ is the number of unique terms in d, $avgdl$ is the average $\mid d \mid$ of all documents, N is the number of documents in the collection, $df(w)$ is the number of documents with w, C is the entire collection, and k_1 and b are constants 2.0 and 0.75.

3.3 Data-driven Approach

To verify the effectiveness of our term weighting schemes in experimental settings of the data-driven approach, we carry out a set of simple experiments with ML classifiers. Specifically, we explore the statistical term weighting features of the word generation model with Support Vector machine (SVM), faithfully reproducing previous work as closely as possible (Pang et al., 2002).

Each instance of train and test data is represented as a vector of features. We test various combinations of the term weighting schemes listed below.

- PRESENCE: binary indicator for the presence of a term
- TF: term frequency
- VS.TF: normalized tf as in VS
- BM25.TF: normalized tf as in BM25
- IDF: inverse document frequency
- VS.IDF: normalized idf as in VS
- BM25.IDF: normalized idf as in BM25

[5]With proper assumptions and derivations, $p(w \mid d)$ can be derived to language modeling approaches. Refer to (Zhai and Lafferty, 2004).

4 Experiment

Our experiments consist of an opinion retrieval task and a sentiment classification task. We use MPQA and movie-review corpora in our experiments with an ML classifier. For the opinion retrieval task, we use the two datasets used by TREC blog track and NTCIR MOAT evaluation workshops.

The opinion retrieval task at TREC Blog Track consists of three subtasks: topic retrieval, opinion retrieval, and polarity retrieval. Opinion and polarity retrieval subtasks use the relevant documents retrieved at the topic retrieval stage. On the other hand, the NTCIR MOAT task aims to find opinionated sentences given a set of documents that are already hand-assessed to be relevant to the topic.

4.1 Opinion Retieval Task – TREC Blog Track

4.1.1 Experimental Setting

TREC Blog Track uses the TREC Blog06 corpus (Macdonald and Ounis, 2006). It is a collection of RSS feeds (38.6 GB), permalink documents (88.8GB), and homepages (28.8GB) crawled on the Internet over an eleven week period from December 2005 to February 2006.

Non-relevant content of blog posts such as HTML tags, advertisement, site description, and menu are removed with an effective internal spam removal algorithm (Nam et al., 2009). While our sentiment analysis model uses the entire relevant portion of the blog posts, further stopword removal and stemming is done for the blog retrieval system.

For the relevance retrieval model, we faithfully reproduce the passage-based language model with pseudo-relevance feedback (Lee et al., 2008).

We use in total 100 topics from TREC 2007 and 2008 blog opinion retrieval tasks (07:901-950 and 08:1001-1050). We use the topics from Blog 07 to optimize the parameter for linearly combining the retrieval and opinion models, and use Blog 08 topics as our test data. Topics are extracted only from the Title field, using the Porter stemmer and a stopword list.

Table 1: Performance of opinion retrieval models using Blog 08 topics. The linear combination parameter λ is optimized on Blog 07 topics. † indicates statistical significance at the 1% level over the baseline.

Model	MAP	R-prec	P@10
TOPIC REL.	0.4052	0.4366	0.6440
BASELINE	0.4141	0.4534	0.6440
VS	0.4196	0.4542	0.6600
BM25	0.4235†	**0.4579**	0.6600
LM	0.4158	0.4520	0.6560
PMI	0.4177	0.4538	0.6620
LSA	0.4155	0.4526	0.6480
WP	0.4165	0.4533	**0.6640**
BM25·PMI	0.4238†	0.4575	0.6600
BM25·LSA	0.4237†	0.4578	0.6600
BM25·WP	0.4237†	**0.4579**	0.6600
BM25·PMI·WP	**0.4242**†	0.4574	0.6620
BM25·LSA·WP	0.4238†	0.4576	0.6580

4.1.2 Experimental Result

Retrieval performances using different combinations of term weighting features are presented in Table 1. Using only the word sentiment model is set as our baseline.

First, each feature of the word generation and topic association models are tested; all features of the models improve over the baseline. We observe that the features of our word generation model is more effective than those of the topic association model. Among the features of the word generation model, the most improvement was achieved with $BM25$, improving the MAP by 2.27%.

Features of the topic association model show only moderate improvements over the baseline. We observe that these features generally improve P@10 performance, indicating that they increase the accuracy of the sentiment analysis system. PMI out-performed LSA for all evaluation measures. Among the topic association models, PMI performs the best in MAP and R-prec, while WP achieved the biggest improvement in P@10.

Since BM25 performs the best among the word generation models, its combination with other features was investigated. Combinations of BM25 with the topic association models all improve the performance of the baseline and BM25. This demonstrates that the word generation model and the topic association model are complementary to each other.

The best MAP was achieved with BM25, PMI, and WP (+2.44% over the baseline). We observe that PMI and WP also complement each other.

4.2 Sentiment Analysis Task – NTCIR MOAT

4.2.1 Experimental Setting

Another set of experiments for our opinion analysis model was carried out on the NTCIR-7 MOAT English corpus. The English opinion corpus for NTCIR MOAT consists of newspaper articles from the Mainichi Daily News, Korea Times, Xinhua News, Hong Kong Standard, and the Straits Times. It is a collection of documents manually assessed for relevance to a set of queries from NTCIR-7 Advanced Cross-lingual Information Access (ACLIA) task. The corpus consists of 167 documents, or 4,711 sentences for 14 test topics. Each sentence is manually tagged with opinionatedness, polarity, and relevance to the topic by three annotators from a pool of six annotators.

For preprocessing, no removal or stemming is performed on the data. Each sentence was processed with the Stanford English parser[6] to produce a dependency parse tree. Only the Title fields of the topics were used.

For performance evaluations of opinion and polarity detection, we use precision, recall, and F-measure, the same measure used to report the official results at the NTCIR MOAT workshop. There are lenient and strict evaluations depending on the agreement of the annotators; if two out of three annotators agreed upon an opinion or polarity annotation then it is used during the lenient evaluation, similarly three out of three agreements are used during the strict evaluation. We present the performances using the lenient evaluation only, for the two evaluations generally do not show much difference in relative performance changes.

Since MOAT is a classification task, we use a threshold parameter to draw a boundary between opinionated and non-opinionated sentences. We report the performance of our system using the NTCIR-7 dataset, where the threshold parameter is optimized using the NTCIR-6 dataset.

4.2.2 Experimental Result

We present the performance of our sentiment analysis system in Table 2. As in the experiments with

[6] http://nlp.stanford.edu/software/lex-parser.shtml

Table 2: Performance of the Sentiment Analysis System on NTCIR7 dataset. System parameters are optimized for F-measure using NTCIR6 dataset with lenient evaluations.

Model	Opinionated		
	Precision	Recall	F-Measure
BASELINE	0.305	**0.866**	0.451
VS	0.331	0.807	**0.470**
BM25	0.327	0.795	0.464
LM	0.325	0.794	0.461
LSA	0.315	0.806	0.453
PMI	0.342	0.603	0.436
DTP	0.322	0.778	0.455
VS·LSA	0.335	0.769	0.466
VS·PMI	0.311	0.833	0.453
VS·DTP	0.342	0.745	0.469
VS·LSA·DTP	**0.349**	0.719	**0.470**
VS·PMI·DTP	0.328	0.773	0.461

the TREC dataset, using only the word sentiment model is used as our baseline.

Similarly to the TREC experiments, the features of the word generation model perform exceptionally better than that of the topic association model. The best performing feature of the word generation model is VS, achieving a 4.21% improvement over the baseline's f-measure. Interestingly, this is the tied top performing f-measure over all combinations of our features.

While LSA and DTP show mild improvements, PMI performed worse than baseline, with higher precision but a drop in recall. DTP was the best performing topic association model.

When combining the best performing feature of the word generation model (VS) with the features of the topic association model, LSA, PMI and DTP all performed worse than or as well as the VS in f-measure evaluation. LSA and DTP improves precision slightly, but with a drop in recall. PMI shows the opposite tendency.

The best performing system was achieved using VS, LSA and DTP at both precision and f-measure evaluations.

4.3 Classification task – SVM

4.3.1 Experimental Setting

To test our SVM classifier, we perform the classification task. Movie Review polarity dataset[7] was first introduced by Pang et al. (2002) to test various ML-based methods for sentiment classification. It is a balanced dataset of 700 positive and 700 negative reviews, collected from the Internet Movie Database (IMDb) archive. MPQA Corpus[8] contains 535 newspaper articles manually annotated at sentence and subsentence level for opinions and other private states (Wiebe et al., 2005).

To closely reproduce the experiment with the best performance carried out in (Pang et al., 2002) using SVM, we use unigram with the **presence** feature. We test various combinations of our features applicable to the task. For evaluation, we use ten-fold cross-validation accuracy.

Table 3: Average ten-fold cross-validation accuracies of polarity classification task with SVM.

Features	Accuracy	
	Movie-review	MPQA
PRESENCE	82.6	76.8
TF	71.1	76.5
VS.TF	81.3	76.7
BM25.TF	81.4	**77.9**
IDF	61.6	61.8
VS.IDF	83.6	**77.9**
BM25.IDF	83.6	77.8
VS.TF·VS.IDF	**83.8**	**77.9**
BM25.TF·BM25.IDF	**84.1**	77.7
BM25.TF·VS.IDF	**85.1**	77.7

4.3.2 Experimental Result

We present the sentiment classification performances in Table 3.

As observed by Pang et al. (2002), using the raw **tf** drops the accuracy of the sentiment classification (-13.92%) of movie-review data. Using the raw **idf** feature worsens the accuracy even more (-25.42%). Normalized **tf**-variants show improvements over **tf** but are worse than **presence**. Normalized **idf** features produce slightly better accuracy results than the baseline. Finally, combining any normalized **tf** and **idf** features improved the baseline (high 83% $\sim$ low 85%). The best combination was **BM25.TF·VS.IDF**.

MPQA corpus reveals similar but somewhat uncertain tendency.

[7] http://www.cs.cornell.edu/people/pabo/movie-review-data/

[8] http://www.cs.pitt.edu/mpqa/databaserelease/

4.4 Discussion

Overall, the opinion retrieval and the sentiment analysis models achieve improvements using our proposed features. Especially, the features of the word generation model improve the overall performances drastically. Its effectiveness is also verified with a data-driven approach; the accuracy of a sentiment classifier trained on a polarity dataset was improved by various combinations of normalized tf and idf statistics.

Differences in effectiveness of VS, BM25, and LM come from parameter tuning and corpus differences. For the TREC dataset, BM25 performed better than the other models, and for the NTCIR dataset, VS performed better.

Our features of the topic association model show mild improvement over the baseline performance in general. PMI and LSA, both modeling the semantic associations between words, show different behaviors on the datasets. For the NTCIR dataset, LSA performed better, while PMI is more effective for the TREC dataset. We believe that the explanation lies in the differences between the topics for each dataset. In general, the NTCIR topics are general descriptive words such as "regenerative medicine", "American economy after the 911 terrorist attacks", and "lawsuit brought against Microsoft for monopolistic practices." The TREC topics are more named-entity-like terms such as "Carmax", "Wikipedia primary source", "Jiffy Lube", "Starbucks", and "Windows Vista." We have experimentally shown that LSA is more suited to finding associations between general terms because its training documents are from a general domain.[9] Our PMI measure utilizes a web search engine, which covers a variety of named entity terms.

Though the features of our topic association model, WP and DTP, were evaluated on different datasets, we try our best to conjecture the differences. WP on TREC dataset shows a small improvement of MAP compared to other topic association features, while the precision is improved the most when this feature is used alone. The DTP feature displays similar behavior with precision. It also achieves the best f-measure over other topic association features. DTP achieves higher relative improvement (3.99% F-measure verse 2.32% MAP), and is more effective for improving the performance in combination with LSA and PMI.

[9]TASA Corpus, http://lsa.colorado.edu/spaces.html

5 Conclusion

In this paper, we proposed various term weighting schemes and how such features are modeled in the sentiment analysis task. Our proposed features include corpus statistics, association measures using semantic and local-context proximities. We have empirically shown the effectiveness of the features with our proposed opinion retrieval and sentiment analysis models.

There exists much room for improvement with further experiments with various term weighting methods and datasets. Such methods include, but by no means limited to, semantic similarities between word pairs using lexical resources such as WordNet (Miller, 1995) and data-driven methods with various topic-dependent term weighting schemes on labeled corpus with topics such as MPQA.

Acknowledgments

This work was supported in part by MKE & IITA through IT Leading R&D Support Project and in part by the BK 21 Project in 2009.

References

Kushal Dave, Steve Lawrence, and David M. Pennock. 2003. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of WWW*, pages 519–528.

Andrea Esuli and Fabrizio Sebastiani. 2006. Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06)*, pages 417–422, Geneva, IT.

Hui Fang, Tao Tao, and ChengXiang Zhai. 2004. A formal study of information retrieval heuristics. In *SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 49–56, New York, NY, USA. ACM.

George Forman. 2003. An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research*, 3:1289–1305.

Michael Gamon. 2004. Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis. In *Proceedings of the International Conference on Computational Linguistics (COLING)*.

Vasileios Hatzivassiloglou and Kathleen R. Mckeown. 1997. Predicting the semantic orientation of adjectives. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL'97)*, pages 174–181, madrid, ES.

Jaap Kamps, Maarten Marx, Robert J. Mokken, and Maarten De Rijke. 2004. Using wordnet to measure semantic orientation of adjectives. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04)*, pages 1115–1118, Lisbon, PT.

Taku Kudo and Yuji Matsumoto. 2004. A boosting algorithm for classification of semi-structured text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Thomas K. Landauer and Susan T. Dumais. 1997. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211–240, April.

Yeha Lee, Seung-Hoon Na, Jungi Kim, Sang-Hyob Nam, Hun young Jung, and Jong-Hyeok Lee. 2008. Kle at trec 2008 blog track: Blog post and feed retrieval. In *Proceedings of TREC-08*.

Craig Macdonald and Iadh Ounis. 2006. The TREC Blogs06 collection: creating and analysing a blog test collection. Technical Report TR-2006-224, Department of Computer Science, University of Glasgow.

Shotaro Matsumoto, Hiroya Takamura, and Manabu Okumura. 2005. Sentiment classification using word subsequences and dependency sub-trees. In *Proceedings of PAKDD'05, the 9th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*.

Qiaozhu Mei, Xu Ling, Matthew Wondra, Hang Su, and ChengXiang Zhai. 2007. Topic sentiment mixture: Modeling facets and opinions in weblogs. In *Proceedings of WWW*, pages 171–180, New York, NY, USA. ACM Press.

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

Tony Mullen and Nigel Collier. 2004. Sentiment analysis using support vector machines with diverse information sources. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 412–418, July. Poster paper.

Sang-Hyob Nam, Seung-Hoon Na, Yeha Lee, and Jong-Hyeok Lee. 2009. Diffpost: Filtering non-relevant content based on content difference between two consecutive blog posts. In *ECIR*.

Vincent Ng, Sajib Dasgupta, and S. M. Niaz Arifin. 2006. Examining the role of linguistic knowledge sources in the automatic identification and classification of reviews. In *Proceedings of the COLING/ACL Main Conference Poster Sessions*, pages 611–618, Sydney, Australia, July. Association for Computational Linguistics.

I. Ounis, M. de Rijke, C. Macdonald, G. A. Mishne, and I. Soboroff. 2006. Overview of the trec-2006 blog track. In *Proceedings of TREC-06*, pages 15–27, November.

I. Ounis, C. Macdonald, and I. Soboroff. 2008. Overview of the trec-2008 blog track. In *Proceedings of TREC-08*, pages 15–27, November.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 79–86.

Fabrizio Sebastiani. 2002. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47.

Yohei Seki, David Kirk Evans, Lun-Wei Ku, Le Sun, Hsin-Hsi Chen, and Noriko Kando. 2008. Overview of multilingual opinion analysis task at ntcir-7. In *Proceedings of The 7th NTCIR Workshop (2007/2008) - Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access*.

Philip J. Stone, Dexter C. Dunphy, Marshall S. Smith, and Daniel M. Ogilvie. 1966. *The General Inquirer: A Computer Approach to Content Analysis*. MIT Press, Cambridge, USA.

Peter D. Turney and Michael L. Littman. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems*, 21(4):315–346.

Peter D. Turney. 2001. Mining the web for synonyms: Pmi-ir versus lsa on toefl. In *EMCL '01: Proceedings of the 12th European Conference on Machine Learning*, pages 491–502, London, UK. Springer-Verlag.

Casey Whitelaw, Navendu Garg, and Shlomo Argamon. 2005. Using appraisal groups for sentiment analysis. In *Proceedings of the 14th ACM international conference on Information and knowledge management (CIKM'05)*, pages 625–631, Bremen, DE.

Janyce Wiebe, E. Breck, Christopher Buckley, Claire Cardie, P. Davis, B. Fraser, Diane Litman, D. Pierce, Ellen Riloff, Theresa Wilson, D. Day, and Mark Maybury. 2003. Recognizing and organizing opinions expressed in the world press. In *Proceedings of the 2003 AAAI Spring Symposium on New Directions in Question Answering*.

Janyce M. Wiebe, Theresa Wilson, Rebecca Bruce, Matthew Bell, and Melanie Martin. 2004. Learning subjective language. *Computational Linguistics*, 30(3):277–308, September.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2/3):164–210.

Theresa Wilson, Janyce Wiebe, and Paul 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 (HLT-EMNLP'05)*, pages 347–354, Vancouver, CA.

Kiduk Yang, Ning Yu, Alejandro Valerio, and Hui Zhang. 2006. WIDIT in TREC-2006 Blog track. In *Proceedings of TREC*.

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

Chengxiang Zhai and John Lafferty. 2004. A study of smoothing methods for language models applied to information retrieval. *ACM Trans. Inf. Syst.*, 22(2):179–214.

Min Zhang and Xingyao Ye. 2008. A generation model to unify topic relevance and lexicon-based sentiment for opinion retrieval. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 411–418, New York, NY, USA. ACM.