Sentiment Retrieval using Generative Models

Koji Eguchi
National Institute of Informatics
Tokyo 101-8430, Japan
eguchi@nii.ac.jp

Victor Lavrenko
Department of Computer Science
University of Massachusetts
Amherst, MA 01003, USA
lavrenko@cs.umass.edu

Abstract

Ranking documents or sentences according to both topic and sentiment relevance should serve a critical function in helping users when topics and sentiment polarities of the targeted text are not explicitly given, as is often the case on the web. In this paper, we propose several sentiment information retrieval models in the framework of probabilistic language models, assuming that a user both inputs query terms expressing a certain topic and also specifies a sentiment polarity of interest in some manner. We combine sentiment relevance models and topic relevance models with model parameters estimated from training data, considering the topic dependence of the sentiment. Our experiments prove that our models are effective.

1 Introduction

The recent rapid expansion of access to information has significantly increased the demands on retrieval or classification of sentiment information from a large amount of textual data. The field of *sentiment classification* has recently received considerable attention, where the polarities of sentiment, such as positive or negative, were identified from unstructured text (Shanahan et al., 2005). A number of studies have investigated sentiment classification at document level, e.g., (Pang et al., 2002; Dave et al., 2003), and at sentence level, e.g., (Hu and Liu, 2004; Kim and Hovy, 2004; Nigam and Hurst, 2005); however, the accuracy is still less than desirable. Therefore, ranking according to the likelihood of containing sentiment information is expected to serve a crucial function in helping users. We believe that our work is the first attempt at *sentiment retrieval* that aims at finding sentences containing information with a specific sentiment polarity on a certain topic.

Intuitively, the expression of sentiment in text is dependent on the topic. For example, a negative view for some voting event may be expressed using 'flaw', while a negative view for some politician may be expressed using 'reckless'. Moreover, sentiment polarities are also dependent on topics or domains. For example, the adjective 'unpredictable' may have a negative orientation in an automotive review, in a phrase such as 'unpredictable steering', but it could have a positive orientation in a movie review, in a phrase such as 'unpredictable plot', as mentioned in (Turney, 2002) in the context of his sentiment word detection.

We propose sentiment retrieval models in the framework of generative language modeling, not only assuming query terms expressing a certain topic, but also assuming that the polarity of sentiment interest is specified by the user in some manner, where the topic dependence of the sentiment is considered. To the best of our knowledge, there have been no other studies on a retrieval model unifying both topic and sentiment, and further, there have been no other studies on sentiment retrieval. The sentiment information often appears as local in a document, and therefore focusing on finer levels, i.e., sentence or passage levels rather than document level, is crucial. We thus experiment on sentiment retrieval at the sentence level in this paper.

The rest of this paper is structured as follows. Section **2** introduces the work related to this study. Section **3** describes a generative model of sentiment, which is proposed here as a theoretical framework for our work. Section **4** describes the task definition and our sentiment retrieval model.

Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006), pages 345–354,
Sydney, July 2006. ©2006 Association for Computational Linguistics

Section **5** explains the data we used for our experiments, and gives our experimental results. Section **6** concludes the paper.

2 Related Work

Some efforts for the TREC Novelty Track were related to our work. Although some of the topics used in the Novelty Track in 2003 and 2004 (Soboroff and Harman, 2003; Soboroff, 2004) were related to opinions, most of the efforts were focused on topic, such as studies using term distribution within each sentence, e.g., (Allan et al., 2003; Losada, 2005; Murdock and Croft, 2005). Amongst the participants in the TREC Novelty Track, only (Kim et al., 2004) proposed a method specialized to opinion-bearing sentence retrieval, by making use of lists of words with positive or negative polarities. They aimed to find opinions on a given topic but did not distinguish or did not care about sentiment polarities that should be represented in some sentences (hereafter, *opinion retrieval*). We focus on finding positive views or negative views according to a given topic and sentiment of interest (hereafter, *sentiment retrieval*). Our work is the first work on sentiment retrieval, to the best of our knowledge.

In the context of *sentiment classification*, some researchers have conducted studies on the topic dependence of sentiment polarities. (Nasukawa and Yi, 2003) and (Yi et al., 2003) extracted positive or negative expressions on a given product name using handmade lexicons. (Engström, 2004) studied how the topic dependence influences the accuracy of sentiment classification and attempted to reduce the influence to improve the accuracy. (Wilson et al., 2005) investigated how context influences sentiment polarity at the phrase level in a corpus, beginning with a predefined list of words with polarities. Their focus on the phenomena of topic dependence of sentiment can be shared with our work; however, their work is not directly related to ours, because we focus on a different task, sentiment retrieval, where different approaches are required.

3 A Generative Model of Sentiment

In this section we will provide a formal underpinning for our approach to sentiment retrieval. The approach is based on the *generative* paradigm: we describe a statistical process that could be viewed, hypothetically, as a source of every statement of interest to our system. We stress that this generative process is to be treated as purely hypothetical; the process is only intended to reflect those aspects of human discourse that are pertinent to the problem of retrieving affectively appropriate and topic-relevant texts in response to a query posed by our user.

Before giving a formal specification of our model, we will provide a high-level overview of the main ideas. We are trying to model a collection of natural-language statements, some of which are relevant to a user's query. In our experiments, these statements are individual sentences, but the model can be applied to textual chunks of any length. We assume that the content of an individual statement can be modeled independently of all other statements in the collection. Each statement consists of some topic-bearing and some sentiment-bearing words. We assume that the topic-bearing words represent exchangeable samples from some underlying topic language model. Exchangeability means that the relative order of the words is irrelevant, but the words are not independent of each other—the idea often stated as a *bag-of-words* assumption. Similarly, sentiment-bearing words are viewed as an order-invariant 'bag', sampled from the underlying sentiment language model. We will explicitly model dependency between the topic and sentiment language models, and will demonstrate that treating them independently leads to sub-optimal retrieval performance. When a *sentiment polarity* value is observed for a given statement, we will treat it as a ternary variable influencing the topic and sentiment language models.

We represent a user's query as just another statement, consisting of topic and sentiment parts, subject to all the independence assumptions stated above. We will use the query to estimate the topic and sentiment language models that are representative of the user's interests. Following (Lavrenko and Croft, 2001), we will use the term *relevance models* to describe these models, and will use them to rank statements in order of their relevance to the query.

3.1 Definitions

We start by providing a set of definitions that will be used in the remainder of this section. The task of our model is to *generate* a collection of statements $\mathbf{w}_1 \ldots \mathbf{w}_n$. A statement $\mathbf{w}_i$ is a string of words $\mathbf{w}_{i1}\ldots\mathbf{w}_{in_i}$, drawn from a common vocabulary $\mathcal{V}$. We introduce a binary variable $b_{ij}{\in}\{S,T\}$ as an indicator of whether the word in the jth position of the ith statement will be a topic word or a sentiment word. For our purposes, b_{ij} is either provided by a human annotator (*manual annotation*), or determined heuristically (*automatic annotation*).

The sentiment polarity x_i for a given statement is a discrete random variable with three outcomes: $\{-1,0,+1\}$, representing negative, neutral and positive polarity values, respectively. As a matter of convenience we will often denote a statement as a triple $\{\mathbf{w}_i^s, \mathbf{w}_i^t, x_i\}$, where $\mathbf{w}_i^s$ contains the sentiment words and $\mathbf{w}_i^t$ contains the topic words. As we mentioned above, the user's query is treated as just another statement. It will be denoted as a triple $\{\mathbf{q}^s, \mathbf{q}^t, \mathbf{q}^x\}$, corresponding to sentiment words, topic keywords, and the desired polarity value. We will use $\mathbf{p}$ to denote a unigram language model, i.e., a function that assigns a number $\mathbf{p}(v){\in}[0,1]$ to every word v in our vocabulary $\mathcal{V}$, such that $\Sigma_v \mathbf{p}(v){=}1$. The set of all possible unigram language models is the probability simplex $\mathbb{P}$. Similarly, $\mathbf{p}_x$ will denote a distribution over the three possible polarity values, and $\mathbb{P}_x$ is the corresponding ternary probability simplex. We define $\pi : \mathbb{P}{\times}\mathbb{P}{\times}\mathbb{P}_x{\rightarrow}[0,1]$ to be a measure function that assigns a probability $\pi(\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_x)$ to a pair of language models $\mathbf{p}_1$ and $\mathbf{p}_2$ together with a polarity model $\mathbf{p}_x$.

3.2 Generative model

Using the definitions presented above, and assuming that $\pi()$ is given, we hypothesize that a new statement $\mathbf{w}_i$ containing words $\mathbf{w}_{i1}\ldots\mathbf{w}_{im}$ with sentiment polarity x_i can be generated according to the following mechanism.

1. Draw $\mathbf{p}_t, \mathbf{p}_s$ and $\mathbf{p}_x$ from $\pi(\cdot,\cdot,\cdot)$.
2. Sample x_i from a polarity distribution $\mathbf{p}_x(\cdot)$.
3. For each position $j = 1\ldots m$:
   (a) if $b_{ij}{=}T$: draw $\mathbf{w}_{ij}$ from $\mathbf{p}_t(\cdot)$;
   (b) if $b_{ij}{=}S$: draw $\mathbf{w}_{ij}$ from $\mathbf{p}_s(\cdot)$.

The probability of observing the new statement $\mathbf{w}_{i1}\ldots\mathbf{w}_{im}$ under this mechanism is given by:

$$\sum_{\mathbf{p}_t,\mathbf{p}_s,\mathbf{p}_x} \pi(\mathbf{p}_t, \mathbf{p}_s, \mathbf{p}_x)\mathbf{p}_x(x_i) \prod_{j=1}^{m} \begin{cases} \mathbf{p}_t(\mathbf{w}_{ij}) & \text{if } b_{ij}{=}T \\ \mathbf{p}_s(\mathbf{w}_{ij}) & \text{otherwise} \end{cases} \tag{1}$$

The summation in equation (1) goes over all possible pairs of language models $\mathbf{p}_t, \mathbf{p}_s$, but we can avoid integration by specifying a mass function $\pi()$ that assigns nonzero probabilities to a finite subset of points in $\mathbb{P}{\times}\mathbb{P}{\times}\mathbb{P}_x$. We accomplish this by using a nonparametric estimate for $\pi()$, the details of which are provided below.

3.2.1 A nonparametric generative mass function

We use a nonparametric estimate for $\pi(\cdot,\cdot,\cdot)$, which makes our generative model similar to *kernel-based* density estimators or *Parzen-window* classifiers (Silverman, 1986). The primary difference is that our model operates over discrete events (strings of words), and accordingly the mass function is defined over the space of distributions, rather than directly over the data points. Our estimate relies on a collection of paired observations $C = \{\mathbf{w}_i^t, \mathbf{w}_i^s, x_i : i{=}1..n\}$, which represent statements for which we know which words are topic words $(\mathbf{w}_i^t)$, and which are sentiment words $(\mathbf{w}_i^s)$. Each of these observations corresponds to a unique point $\mathbf{p}_{ti}, \mathbf{p}_{si}, \mathbf{p}_{xi}$ in the space of paired distributions $\mathbb{P}{\times}\mathbb{P}{\times}\mathbb{P}_x$, defined by the following coordinates:

$$\begin{aligned} \mathbf{p}_{ti}(v) &= \lambda_t \#(v, \mathbf{w}_i^t)/\#(\mathbf{w}_i^t) + (1{-}\lambda_t)c_{tv} \\ \mathbf{p}_{si}(v) &= \lambda_s \#(v, \mathbf{w}_i^s)/\#(\mathbf{w}_i^s) + (1{-}\lambda_s)c_{sv} \\ \mathbf{p}_{xi}(x) &= \lambda_x 1_{x=x_i} + (1{-}\lambda_x). \end{aligned} \tag{2}$$

Here, $\#(v, \mathbf{w}_i^t)$ represents the number of times the word v was observed in the topic part of statement i, the length of which is denoted by $\#(\mathbf{w}_i^t)$. c_{tv} stands for the relative frequency of v in the topic part of the collection. The same definitions apply to the sentiment parameters $\#(v, \mathbf{w}_i^s)$, $\#(\mathbf{w}_i^s)$ and c_{sv}. The Boolean indicator function 1_y returns one when the predicate y is true and zero otherwise. Metaparameters λ_t, λ_s and λ_x specify the amount of Dirichlet smoothing (Zhai and Lafferty, 2001) applied to the topic, sentiment and polarity estimates respectively; values for these parameters are determined empirically.

We define $\pi(\mathbf{p}_t, \mathbf{p}_s, \mathbf{p}_x)$ to have mass $\frac{1}{n}$ when its argument $\mathbf{p}_t, \mathbf{p}_s, \mathbf{p}_x$ corresponds to some observation $\mathbf{p}_{ti}, \mathbf{p}_{si}, \mathbf{p}_{xi}$, and zero otherwise:

$$\pi(\mathbf{p}_t, \mathbf{p}_s, \mathbf{p}_x) = \frac{1}{n}\sum_{i=1}^{n} 1_{\mathbf{p}_t=\mathbf{p}_{ti}} {\times} 1_{\mathbf{p}_s=\mathbf{p}_{si}} {\times} 1_{\mathbf{p}_x=\mathbf{p}_{xi}}. \tag{3}$$

Equation (3) maintains empirical dependencies between the topic language model $\mathbf{p}_t$ and the sentiment model $\mathbf{p}_s$, because we assign nonzero probability mass only to pairs of models that actually *co-occur* in our observations.

3.2.2 Limitations of the model

Our model represents each statement $\mathbf{w}_i$ as a *bag of words*, or more formally an order-invariant sequence. This representation is often confused with *word independence*, which is a much stronger assumption. The generative model defined by equation (1) ignores the relative ordering of the words, but it does allow arbitrarily strong *unordered dependencies* among them. To illustrate, consider the probability of observing the words ‘unpredictable’ and ‘plot’ in the same statement. Suppose we set $\lambda_t, \lambda_s{=}1$ in equation (2), reducing the effects of smoothing. It should be evident that $P(\text{unpredictable,plot})$ will be non-zero only when the two words actually co-occur in the training data. By carefully selecting the smoothing parameters, the model can preserve dependencies between topic and sentiment words, and is quite capable of distinguishing the positive sentiment of ‘unpredictable plot’ from the negative sentiment of ‘unpredictable steering’. On the other hand, the model does ignore the ordering of the words, so it will not be able to differentiate the negative phrase ‘gone from good to bad’ from its exact opposite. Furthermore, our model is not well suited for modeling adjacency effects: the phrase ‘unpredictable plot’ is treated in the same way as two separate words, ‘unpredictable’ and ‘plot’, co-occurring in the same sentence.

3.3 Using the model for retrieval

The generative model presented above can be applied to sentiment retrieval in the following fashion. We start with a collection of statements C and a query $\{\mathbf{q}^s, \mathbf{q}^t, \mathbf{q}^x\}$ supplied by the user. We use the machinery outlined in Section **3.2** to estimate the topic and sentiment relevance models corresponding to the user’s information need, and then determine which statements in our collection most closely correspond to these models of relevance. The topic relevance model R_t and sentiment relevance model R_s are estimated as follows. We assume that our query $\mathbf{q}^s, \mathbf{q}^t, \mathbf{q}^x$ is a random sample from a distribution defined by equation (1), and then for each word v we estimate the likelihood that v would be observed if we sampled one more topic or sentiment word:

$$R_t(v){=}\frac{P(\mathbf{q}^s, \mathbf{q}^t{\circ}v, \mathbf{q}^x)}{P(\mathbf{q}^s, \mathbf{q}^t, \mathbf{q}^x)},\ R_s(v){=}\frac{P(\mathbf{q}^s{\circ}v, \mathbf{q}^t, \mathbf{q}^x)}{P(\mathbf{q}^s, \mathbf{q}^t, \mathbf{q}^x)}. \tag{4}$$

Both the numerator and denominator are computed according to equation (1), with the mass function $\pi()$ given by equations (3) and (2). We use the notation $\mathbf{q}{\circ}v$ to denote appending word v to the string $\mathbf{q}$. Estimation is done over the training corpus, which may or may not include numeric values of sentiment polarity.[1] Once we have estimates for the topic and sentiment relevance models, we can rank testing statements $\mathbf{w}$ by their similarity to R_t and R_s. We rank statements using a variation of cross-entropy, which was proposed by (Zhai, 2002):

$$\alpha \sum_v R_t(v) \log \mathbf{p}_t(v){+}(1{-}\alpha) \sum_v R_s(v) \log \mathbf{p}_s(v). \tag{5}$$

Here the summations extend over all words v in the vocabulary, R_t and R_s are given by equation (4), while $\mathbf{p}_t$ and $\mathbf{p}_s$ are computed according to equation (2). A weighting parameter α allows us to change the balance of topic and sentiment in the final ranking formula; its value is selected empirically.

4 Sentiment Retrieval Task

4.1 Task definition

We define two variations of the sentiment retrieval task. In one, the user supplies us with a numeric value for the desired polarity $\mathbf{q}^x$. In the other, the user supplies a set of *seed words* $\mathbf{q}^s$, reflecting the desired sentiment. The first task requires us to have polarity observations x_i in our training data, while the second does not.

Task with training data:

Input: (1) a set of topic keywords $\mathbf{q}^t$ and (2) a sentiment specification $\mathbf{q}^x \in \{-1, 1\}$. In this case we assume $\mathbf{q}^s$ to be the empty string.

Output: a ranked list of topic-relevant and sentiment-relevant sentences from the test data.

Task with seed words:

Input: (1) a set of topic keywords $\mathbf{q}^t$ and (2) a set of sentiment seed words $\mathbf{q}^s$. In this case our model ignores $\mathbf{q}^x$ and x_i.

[1] When the training corpus does not contain numeric polarity values x_i, we assume $\pi(\mathbf{p}_t, \mathbf{p}_s, \mathbf{p}_x){=}\pi(\mathbf{p}_t, \mathbf{p}_s)$ and force $\mathbf{p}_x(x_i)$ to be a constant.

Output: a ranked list of topic-relevant and sentiment-relevant sentences from the test data.

In the first task, we split our corpus into three parts: (i) the training set, which was used for estimating the relevance models R_s and R_t; (ii) the development set, which was used for tuning the model parameters λ_t, λ_s and α; and (iii) the testing set, from which we retrieved sentences in response to the query. In the second task, we split the corpus into two parts: (i) the training set, which was used for tuning the model parameters; and (ii) the testing set, which was used for constructing R_s and R_t and from which we retrieved sentences in response to queries.[2] The testing set was identical in both tasks. Note that the sentiment relevance model R_s can be constructed in a topic-dependent fashion for both tasks.

4.2 Variations of the retrieval model

slm: the retrieval model as described in Section **3.3**.

lmt: the standard language modeling approach (Ponte and Croft, 1998; Song and Croft, 1999) on the topic keywords $\mathbf{q}^t$ for the topic part of the text $\mathbf{w}^t$.

lms: the standard language modeling approach on the sentiment keywords $\mathbf{q}^s$ for the sentiment part of the text $\mathbf{w}^s$.

base: the weighted linear combination of *lmt* and *lms*.

rmt: only the topic relevance model was used for ranking using $\mathbf{q}^t$ and for $\mathbf{w}^t$.[3]

rms: only the sentiment relevance model was used for ranking using $\mathbf{q}^s$ and for $\mathbf{w}^s$.

rmt-base: the *slm* model with $\alpha = 1$, ignoring the sentiment relevance model.

rms-base: the *slm* model with $\alpha = 0$, ignoring the topic relevance model.

rmt-rms: the *rmt* and *rms* models are treated independently.

rmt-slm: the *rmt* and *rms-base* models are combined.

lmtf: the standard language modeling approach using $\mathbf{q}^t$ for the nonsplit text, as baseline.

rmtf: the conventional relevance model was used for ranking using $\mathbf{q}^t$ for the nonsplit text, as baseline.

lmtsf: the standard language modeling approach using both $\mathbf{q}^t$ and $\mathbf{q}^s$ for the nonsplit text, for reference.

rmtsf: the conventional relevance model was used for ranking using both $\mathbf{q}^t$ and $\mathbf{q}^s$ for the nonsplit text, for reference.

Note that the relevance models are constructed using training data for the training-based task, but are constructed using test data for the seed-based task, as mentioned in Section **4.1**. Therefore, the *base* model is only used for the training data, not for the test data, in the training-based task, while it can be performed for the test data in the case of the seed-based task. Moreover, the *lms*, *lmtsf* and *rmtsf* models are based on the premise of using seed words to specify sentiments, and so they are only applicable to the seed-based task.

In the models described in this subsection, λ_t and λ_s in equation (2) were set to Dirichlet estimates (Zhai and Lafferty, 2001), $\#(\mathbf{w}_i^t)/(\#(\mathbf{w}_i^t) + \mu_t)$ and $\#(\mathbf{w}_i^s)/(\#(\mathbf{w}_i^s) + \mu_s)$ for the relevance models R_t and R_s, respectively, in equation (4), and were fixed at 0.9 for ranking as in equation (5) for our experiments in Section **5**. Here, μ_t and μ_s were selected empirically according to the tasks described in Section **4.1**. The model parameter α in equation (5) was also selected empirically in the same manner. The number of ranked documents used in the relevance models R_t and R_s, in equation (4), was selected empirically in the same manner as above; however, we fixed the number of terms used in the relevance models as 1000.

5 Experiments

5.1 Data set and evaluation measure

We used the MPQA Opinion Corpus version 1.2 (Wilson et al., 2005; Wiebe et al., 2005) to measure the effectiveness of our sentiment re-

[2] Because the training set was used for tuning the model parameters, no development set was required for this task.

[3] When we use the automatic annotation that is described in Section **5.2.2**, we use the whole text instead of the topic part of the text, for the reasons given in that section. This treatment is applied to the *base*, *rmt-base*, *rms-base*, *rmt-rms*, *rmt-slm* and *slm* models that are described in this section for using the automatic annotation. However, we distinguish the *lmt* and *rmt* models using the topic part of the text and the *lmtf* and *rmtf* models, as baselines, using the whole text, respectively, even in the experiments using the automatic annotation.

trieval models. We summarize this data set as follows.

- This corpus contains news articles collected from 187 different foreign and U.S. news sources from June 2001 to May 2002. The corpus contains 535 documents, a total of 11,114 sentences.
- The majority of the articles are on 10 different topics, which are labeled at document level, but, in addition to these, a number of additional articles were randomly selected from a larger corpus of 270,000 documents.
- Each article was manually annotated using an annotation scheme for opinions and other private states at phrase level. We only used the annotations for sentiments that included some attributes such as polarity and strength.

In this data set, the topic relevance for the 10 topics is known at the document level, but unknown at the sentence level. We assumed that all the sentences in a relevant document could be considered relevant to the topic.[4]

This data set was annotated with sentiment polarities at the phrase level, but not explicitly annotated at the sentence level. Therefore, we provided sentiment polarities at the sentence level to prepare training data and data for evaluation. We set the sentence-level sentiment polarity equal to the polarity with the highest strength in each sentence.[5]

Queries were expressed using the title of one of the 10 topics and specified as positive or negative. Thus, we had 20 types of queries for our experiments. Because the supposed relevance judgments in this setting are imperfect at sentence level, we used *bpref* (Buckley and Voorhees, 2004), in both the training and testing phases, as it is known to be tolerant of imperfect judgments. Bpref uses binary relevance judgments to define the preference relation (i.e., any relevant document is preferred over any nonrelevant document for a given topic), while other measures, such as mean average precision, depend only on the ranks of the relevant documents.

5.2 Extracting sentiment expressions

5.2.1 Using manual annotation

Because the MPQA corpus was annotated with phrase-level sentiments, we can use these annotations to split a sentence into a topic part $\mathbf{w}^t$ and a sentiment part $\mathbf{w}^s$. The Krovetz stemmer (Krovetz, 1993) was applied to the topic part, the sentiment part and to the query terms[6] and, for the retrieval experiments in Sections **5.3** and **5.4**, a total of 418 stopwords from a standard stopword list were removed when they appeared.

5.2.2 Using automatic annotation

In automatic extraction of sentiment expressions in this study, we detected sentiment-bearing words using lists of words with established polarities. At this stage, topic dependence was not considered; however, at the stage of sentiment modeling, the topic dependence can be reflected, as described in Sections **3** and **4**.

We first prepared a list of words indicating sentiments. We used Hatzivassiloglou and McKeown's sentiment word list (Hatzivassiloglou and McKeown, 1997), which consists of 657 positive and 679 negative adjectives, and The General Inquirer (Stone et al., 1966), which contains 1621 positive and 1989 negative words.[7] By merging these lists, we obtained 1947 positive and 2348 negative words. After stemming these words in the same manner as in Section **5.2.1**, we were left with 1667 positive and 2129 negative words, which we will use hereafter in this paper.

The sentiment polarities are sometimes sensitive to the structural information, for instance, a negation expression reverses the following sentiment polarity. To handle negation, every sentiment-bearing word was rewritten with a 'NEG' suffix, such as 'good_NEG', if an odd number of negation expressions was found within the five preceding words in the sentence. To detect negation expressions, we used a predefined negation expression list. This negation handling is similar to that used in (Das and Chen, 2001; Pang et al., 2002). We extracted sentiment-bearing expressions using the list of words with established po-

[4] This is a strong assumption to make and may not be true in all cases. A larger, more complete data set is required to perform a more detailed analysis, which is left as future work.

[5] We disregarded 'neutral' and 'both' if other polarities appeared. We can also set the sentence-level sentiment polarity according to the presence of polarity in each sentence, but we did not consider this setting here.

[6] We used the topic labels attached to the MPQA corpus as the topic query terms $\mathbf{q}^t$ in all the experiments in Sections **5.3** and **5.4**.

[7] We extracted positive and negative words from the General Inquirer basically in the same manner as in (Turney and Littman, 2003); however, we did not exclude any words, unlike (Turney and Littman, 2003), where some seed words were excluded for the evaluation of their work.

Table 1: Sample probabilities from the sentiment relevance models

Topic-independent		Topic-independent		Reaction to President Bush's 2002 State of the Union Address				2002 presidential election in Zimbabwe				Israeli settlements in Gaza and West Bank			
w/ manual annot.		w/ automatic annot.		w/ manual annot.		w/ automatic annot.		w/ manual annot.		w/ automatic annot.		w/ manual annot.		w/ automatic annot.	
$P(w\|Q)$	w	$P(w\|Q)$	w	$P(w\|Q)$	w	$P(w\|Q)$	w	$P(w\|Q)$	w	$P(w\|Q)$	w	$P(w\|Q)$	w	$P(w\|Q)$	w
0.047	demand	0.029	state	0.030	support	0.067	state	0.042	support	0.039	support	0.041	ask	0.097	settle
0.031	expect	0.026	support	0.016	promise	0.034	support	0.033	legitimate	0.033	legitimate	0.036	agreed	0.032	peace
0.031	defend	0.014	lead	0.014	call	0.024	call	0.031	free	0.033	lead	0.036	call	0.025	state
0.031	invite	0.013	call	0.014	excellent	0.019	meet	0.029	congratulate	0.025	free	0.033	aim	0.022	secure
0.031	humane	0.013	minister	0.013	goal	0.017	minister	0.028	fair	0.025	fair	0.028	immediate	0.015	call
0.031	safeguard	0.011	right	0.013	express	0.015	promise	0.023	please	0.018	state	0.025	aware	0.014	conflict
0.031	nutritious	0.010	foreign	0.013	best	0.014	white	0.017	confident	0.017	congratulate	0.024	key	0.013	support
0.031	helpful	0.009	hope	0.012	count	0.013	foreign	0.017	call	0.015	call	0.022	expect	0.012	right
0.016	time	0.009	meet	0.012	cooperate	0.012	success	0.012	hopeful	0.015	meet	0.018	justify	0.011	attack
0.016	say	0.008	interest	0.011	proposal	0.011	defense	0.012	express	0.013	unity	0.018	honoure	0.011	minister
0.091	evil	0.037	state	0.065	evil	0.098	state	0.029	flaw	0.028	flaw	0.018	palestinian	0.100	settle
0.080	axis	0.022	evil	0.049	axis	0.051	evil	0.018	condemn	0.026	critic	0.013	protest	0.031	state
0.045	threat	0.015	right	0.022	critic	0.028	critic	0.015	true	0.023	state	0.012	decide	0.019	peace
0.033	qualify	0.015	prison	0.011	prepare	0.017	call	0.014	critic	0.022	opposition	0.011	peace	0.014	secure_NEG
0.030	wrote	0.013	critic	0.010	recognize	0.012	interest	0.012	expect	0.019	reject	0.011	fatten	0.013	critic
0.020	particular	0.010	human	0.010	reckless	0.011	move	0.011	reject	0.017	condemn	0.011	believe	0.012	force
0.020	word	0.008	support	0.010	country	0.011	reject	0.011	s	0.016	legal	0.009	plan	0.012	attack
0.018	harsh	0.008	protest	0.009	upset	0.010	slam	0.011	fair	0.015	move	0.009	fear	0.012	war
0.015	reject	0.008	war	0.009	pick	0.010	right	0.011	free	0.015	democratic	0.009	mistake	0.011	believe
0.015	dangerous	0.008	force	0.009	eyesore	0.010	attack	0.010	angry	0.014	support	0.009	continue	0.011	minister

The upper and lower tables correspond to positive and negative sentiments, respectively. The topic-independent sentiment relevance models (in the left two columns) correspond to *rms*, and the topic-dependent models (in the rest of the columns) correspond to *rms-base*, which is used for *slm*.

larities, considering negation, as described above. Note that we used the list of words with sentiments to extract sentiment expressions, but we did not use the predefined sentiments to model sentiment relevance.

Some expressions are sometimes used to express a certain topic, such as *settlements* in "Israeli *settlements* in Gaza and West Bank"; but at other times are used to express a certain sentiment, such as the same word in "All parties signed court-mediated compromise *settlements*". Therefore, we will use whole sentences to model topic relevance, while we will use the automatically extracted sentiment expressions to model sentiment relevance, in Sections **5.3** and **5.4**.

5.3 Experiments on training-based task

We conducted experiments on the training-based task described in Section **4.1**, using either manual annotation as described in Section **5.2.1** or automatic annotation as described in Section **5.2.2**. **Table 1** contrasts sample probabilities from topic-independent sentiment relevance models and those from topic-dependent sentiment relevance models. In the left two columns of this table, two sets of sample probabilities using the topic-independent model are presented. One was computed from the manual annotation and the other was computed from the automatic annotation. In the remaining columns, samples using the topic-dependent model are shown according to the three topics: (1) "reaction to President Bush's 2002 State of the Union Address", (2) "2002 presidential election in Zimbabwe", and (3) "Israeli settlements in Gaza and West Bank". A number of positive expressions appeared topic dependent, such as 'promise' (stemmed from 'promising' or not) and 'support' for Topic (1), 'legitimate' and 'congratulate' for Topic (2) and 'justify' and 'secure' for Topic (3); while negative expressions appeared topic-dependent, such as 'critic' (stemmed from 'criticism') and 'eyesore' for Topic (1), 'flaw' and 'condemn' for Topic (2) and 'mistake' and 'secure_NEG' (i.e., 'secure' was negated) for Topic (3).

Some expressions were unexpectedly generated regardless of the types of annotation, e.g., 'palestinian' for Topic (3); however, we found some characteristics in the results using automatic annotation. Some expressions on opinions that did not convey sentiments, such as 'state', frequently appeared regardless of topic. This sort of expression may effectively function as degrading sentences only conveying facts, but may function harmfully by catching sentences conveying opinions without sentiments in the task of sentiment retrieval. Some topic expressions, such as 'settle' (stemmed from 'settlement' or not) for Topic (3), were generated, because such words convey positive sentiments in some other contexts and thus they were contained in the list of sentiment-bearing words that we used for automatic annotation. This will not cause a topic relevance model to drift, because we modeled the topic relevance using whole sentences, as described in Section **5.2.2**; however, it may harm the sentiment relevance model to some extent.

Table 2: Experimental results of training-based task using manually annotated data

	10%		25%		40%	
Models	Bpref	(AvgP)	Bpref	(AvgP)	Bpref	(AvgP)
lmtf	0.1389	(0.1135)	0.1389	(0.1135)	0.1386	(0.1145)
lmt	0.1499	(0.1164)	0.1499	(0.1164)	0.1444	(0.1148)
rmtf	0.1811	(0.1706)	0.1887	(0.1770)	0.1841	(0.1691)
rmt	0.1712	(0.1619)	0.1712	(0.1619)	0.1922	(0.1705)
rmt-base	0.1922	(0.1723)	0.2005	(0.1812)	0.2100*	(0.1951)
rms	0.0464	(0.0384)	0.0452	(0.0394)	0.0375	(0.0320)
rms-base	0.0772	(0.0640)	0.0869	(0.0704)	0.0865	(0.0724)
rmt-rms	0.2025	(0.1413)	0.2210	(0.1925)	0.2117	(0.2003)
rmt-slm	0.2278*	(0.1715)	0.2249	(0.1676)	0.1999	(0.1819)
slm	0.2006	(0.1914)	0.2247	(0.1824)	0.2441*	(0.2427)

‘*’ indicates statistically significant improvement over *rmtf* where $p < 0.05$ with the two-sided Wilcoxon signed-rank test.

Table 3: Experimental results of training-based task using automatically annotated data

	10%		25%		40%	
Models	Bpref	(AvgP)	Bpref	(AvgP)	Bpref	(AvgP)
lmtf	0.1389	(0.1135)	0.1389	(0.1135)	0.1386	(0.1145)
lmt	0.1325	(0.0972)	0.1315	(0.0976)	0.1325	(0.0972)
rmtf	0.1811	(0.1706)	0.1887	(0.1770)	0.1841	(0.1691)
rmt	0.1490	(0.1418)	0.1762	(0.1584)	0.1695	(0.1485)
rmt-base	0.2076*	(0.1936)	0.2252*	(0.2139)	0.2302*	(0.2196)
rms	0.0347	(0.0287)	0.0501	(0.0408)	0.0501	(0.0408)
rms-base	0.0943	(0.0733)	0.1196	(0.0896)	0.1241	(0.0979)
rmt-rms	0.1690	(0.1182)	0.2063	(0.1938)	0.1603	(0.1591)
rmt-slm	0.1980	(0.1426)	0.2013	(0.1835)	0.2148	(0.1882)
slm	0.2011	(0.1537)	0.2261*	(0.1716)	0.2318*	(0.1802)

‘*’ indicates statistically significant improvement over *rmtf* where $p < 0.05$ with the two-sided Wilcoxon signed-rank test.

We performed retrieval experiments in the steps described in Section **4.1**. For this purpose, we split the data into three parts: (i) x% as the training data, (ii) $(50 - x)$% as the evaluation data, and (iii) 50% as the test data.

The test results of training-based task using manually annotated data and automatically annotated data are shown in **Tables 2** and **3**, respectively. The scores were computed according to the *bpref* evaluation measure (Buckley and Voorhees, 2004), as mentioned in Section **5.1**. In addition to the bpref, mean average precision values are presented as ‘AvgP’ in the tables, for reference.[8] In these tables, the top row indicates the percentages of the training data x. It turned out that in all our experiments the appropriate fraction of training data was 40%. In this setting, our *slm* model worked 76.1% better than the query likelihood model and 32.6% better than the conventional relevance model, when using manual annotation, and both improvements were statistically significant according to the Wilcoxon signed-rank test.[9] When using automatic annotation, the *slm* model worked 67.2% better than the query likelihood model and 25.9% better than the conventional relevance model, where both improvements were statistically significant. The *rmt-base* model also worked well with automatic annotation.

5.4 Experiments on seed-based task

For experiments on the seed-based task that was described in Section **4.1**, we used three groups of seed words: KAM, TUR and ORG. Each group consists of a positive word set $\mathbf{q}^s_{(+)}$ and a negative word set $\mathbf{q}^s_{(-)}$, as follows:

KAM: $\mathbf{q}^s_{(+)} = \{$good$\}$, and $\mathbf{q}^s_{(-)} = \{$bad$\}$.

TUR: $\mathbf{q}^s_{(+)} = \{$good, nice, excellent, positive, fortunate, correct, superior$\}$, and $\mathbf{q}^s_{(-)} = \{$bad, nasty, poor, negative, unfortunate, wrong, inferior$\}$.

ORG: $\mathbf{q}^s_{(+)} = \{$support, demand, promise, want, hope$\}$, and $\mathbf{q}^s_{(-)} = \{$refuse, accuse, criticism, fear, reject$\}$.

KAM and TUR were used in (Kamps and Marx, 2002) and (Turney and Littman, 2003), respectively. We constructed ORG considering sentiment-bearing words that may frequently appear in newspaper articles.

We experimented with the seed-based task, making use of each of these seed word groups, in the steps described in Section **4.1**. For this purpose, we split the data into two parts: (i) 50% as the estimation data and (ii) 50% as the test data.

The test results using manually annotated data and automatically annotated data are shown in **Tables 4** and **5**, respectively, where the scores were computed according to the bpref evaluation measure. Mean average precision values are also presented as ‘AvgP’ in the tables, for reference.

When using the manually annotated approach, our *slm* model worked well, especially with the seed word group ORG, as shown in **Table 4**. Using ORG, the *slm* model worked 61.2% better than the query likelihood model and 15.2% better than the conventional relevance model, where both improvements were statistically significant according to the Wilcoxon signed-rank test. Even

[8] As mentioned in Section **5.1**, the bpref is more appropriate for the evaluation of our experiments than the mean average precision.

[9] Significance tests involved only 20 queries, which makes it difficult to achieve statistical significance.

Table 4: Experimental results of seed-based task using manually annotated data

	ORG		TUR		KAM	
Models	Bpref	(AvgP)	Bpref	(AvgP)	Bpref	(AvgP)
lmtf	0.1385	(0.1119)	0.1385	(0.1119)	0.1385	(0.1119)
lmtsf	0.1182	(0.1035)	0.1061	(0.0884)	0.1330	(0.1062)
lmt	0.1501	(0.1171)	0.1501	(0.1171)	0.1501	(0.1171)
base	0.1615	(0.1319)	0.1531	(0.1217)	0.1514	(0.1180)
rmtf	0.1938	(0.1776)	0.1938	(0.1776)	0.1938	(0.1776)
rmtsf	0.1884	(0.1775)	0.1661	(0.1412)	0.1927	(0.1754)
rmt	0.1974	(0.1826)	0.1974	(0.1826)	0.1974	(0.1826)
rmt-base	0.1960	(0.1918)	0.1931	(0.1703)	0.1837	(0.1721)
rms	0.0434	(0.0262)	0.0295	(0.0205)	0.0280	(0.0170)
rms-base	0.1142	(0.1022)	0.1144	(0.0841)	0.1226	(0.0973)
rmt-rms	0.1705	(0.1117)	0.1403	(0.1424)	0.1405	(0.0842)
rmt-slm	0.2266*	(0.2034)	0.2272*	(0.2012)	0.2264*	(0.2016)
slm	0.2233*	(0.2048)	0.2160	(0.1945)	0.2072	(0.1929)

'*' indicates statistically significant improvement over *rmtf* where $p < 0.05$ with the two-sided Wilcoxon signed-rank test.

Table 5: Experimental results of seed-based task using automatically annotated data

	ORG		TUR		KAM	
Models	Bpref	(AvgP)	Bpref	(AvgP)	Bpref	(AvgP)
lmtf	0.1385	(0.1119)	0.1385	(0.1119)	0.1385	(0.1119)
lmtsf	0.1182	(0.1035)	0.1061	(0.0884)	0.1330	(0.1062)
lmt	0.1325	(0.0972)	0.1325	(0.0972)	0.1325	(0.0972)
basef	0.1550	(0.1369)	0.1451	(0.1188)	0.1416	(0.1142)
rmtf	0.1938	(0.1776)	0.1938	(0.1776)	0.1938	(0.1776)
rmtsf	0.1884	(0.1775)	0.1661	(0.1412)	0.1927	(0.1754)
rmt	0.1757	(0.1578)	0.1757	(0.1578)	0.1757	(0.1578)
rmt-base	0.1957	(0.1862)	0.1976	(0.1882)	0.1825	(0.1704)
rms	0.0421	(0.0236)	0.0364	(0.0205)	0.0217	(0.0147)
rms-base	0.1268	(0.1096)	0.1301	(0.1148)	0.1326	(0.1158)
rmt-rms	0.1465	(0.1514)	0.1390	(0.1393)	0.1252	(0.0757)
rmt-slm	0.1977	(0.1811)	0.2008	(0.1649)	0.1959	(0.1677)
slm	0.2031	(0.1714)	0.2055*	(0.1668)	0.2044*	(0.1698)

'*' indicates statistically significant improvement over *rmtf* where $p < 0.05$ with the two-sided Wilcoxon signed-rank test.

using the other seed word groups, the *slm* model worked 49–56% better than the query likelihood model and 6–12% better than the conventional relevance model; however, the latter improvement was not statistically significant. The *rmt-slm* model also worked well with manual annotation.

When using automatic annotation, the *slm* model worked 46–48% better than the query likelihood model and 4–6% better than the conventional relevance model, as shown in **Table 5**. The improvements over the conventional relevance model were statistically significant only when using TUR or KAM; however, the score when using ORG is almost comparable with the others.

6 Conclusion

We propose sentiment retrieval models in the framework of probabilistic generative models, not only assuming that a user inputs query terms expressing a certain topic, but also assuming that the user specifies a sentiment polarity of interest either as a sentiment specification $\mathbf{q}^x \in \{-1, 1\}$ or as a set of sentiment seed words $\mathbf{q}^s$. For this purpose, we combine sentiment relevance models and topic relevance models, considering the topic dependence of the sentiment. In our experiments, our model worked significantly better than standard language modeling approaches, both when using $\mathbf{q}^x$ and $\mathbf{q}^s$, and with both manual and automatic annotation of the fragments expressing sentiments in text. With $\mathbf{q}^s$ and automatic annotation, our model still worked significantly better than the standard approaches; however, the performance did not reach that achieved with other settings. We believe the performance can be improved with larger-scale data.

We experimented to find sentences that were relevant to a given topic and were appropriate to a given sentiment; however, our models can also be applied to textual chunks of any length, such as at document level or passage level. Our model can be easily extended to *opinion retrieval*, if the opinion retrieval is defined as retrieving sentences or documents that contain either positive or negative sentiments. This issue is worth pursuing in future work. Approaches considering polarity strength or continuous values for the polarity specification, rather than using $\{-1, 1\}$, can also be considered in future work.

Acknowledgments

We thank James Allan, W. Bruce Croft and the anonymous reviewers for valuable discussions and comments. This work was supported in part by the Overseas Research Scholars Program and the Grant-in-Aid for Scientific Research (#17680011) from the Ministry of Education, Culture, Sports, Science and Technology, Japan, in part by the Telecommunications Advancement Foundation, Japan, in part by the Center for Intelligent Information Retrieval, and in part by the Defense Advanced Research Projects Agency (DARPA), USA under contract number HR0011-06-C-0023. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect those of the sponsor.

References

James Allan, Courtney Wade, and Alvaro Bolivar. 2003. Retrieval and novelty detection at the sentence level. In *Proc. of the 26th Annual International ACM SIGIR Conference*, pages 314–321, Toronto, Canada.

Chris Buckley and Ellen M. Voorhees. 2004. Retrieval evaluation with incomplete information. In *Proc. of the 27th Annual International ACM SIGIR Conference*, pages 25–32, Sheffield, United Kingdom.

Sanjiv R. Das and Mike Y. Chen. 2001. Yahoo! for Amazon: Sentiment parsing from small talk on the Web. In *Proc. of the 2001 European Finance Association Annual Conference*, Barcelona, Spain.

Kushal Dave, Steve Lawrence, and David M. Pennock. 2003. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proc. of the 12th International Conference on the World Wide Web*, pages 519–528, Budapest, Hungary.

Charlotta Engström. 2004. Topic dependence in sentiment classification. Master's thesis, University of Cambridge.

Vasileios Hatzivassiloglou and Kathleen R. McKeown. 1997. Predicting the semantic orientation of adjectives. In *Proc. of the 35th Annual Meeting of the Association for Computational Linguistics*, pages 174–181, Madrid, Spain.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proc. of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 168–177, Seattle, USA.

Jaap Kamps and Maarten Marx. 2002. Words with attitude. In *Proc. of the 1st International Conference on Global WordNet*, pages 332–341, Mysore, India.

Soo-Min Kim and Eduard Hovy. 2004. Determining the sentiment of opinions. In *Proc. of the 20th International Conference on Computational Linguistics*, Geneva, Czech Republic.

Soo-Min Kim, Deepak Ravichandran, and Eduard Hovy. 2004. ISI Novelty Track system for TREC 2004. In *Proc. of the 13th Text Retrieval Conference*. NIST Special Publication 500-261.

Robert Krovetz. 1993. Viewing morphology as an inference process. In *Proc. of the 16th Annual International ACM SIGIR Conference*, pages 191–202, Pittsburgh, Pennsylvania, USA.

Victor Lavrenko and W. Bruce Croft. 2001. Relevance-based language models. In *Proc. of the 24th Annual International ACM-SIGIR Conference*, pages 120–127, New Orleans, Louisiana, USA.

David E. Losada. 2005. Language modeling for sentence retrieval: A comparison between multiple-Bernoulli and multinomial models. In *Information Retrieval and Theory Workshop*, Glasgow, United Kingdom.

Vanessa Murdock and W. Bruce Croft. 2005. A translation model for sentence retrieval. In *Proc. of HLT/EMNLP 2005*, pages 684–691, Vancouver, Canada.

Tetsuya Nasukawa and Jeonghee Yi. 2003. Sentiment analysis: Capturing favorability using natural language processing. In *Proc. of the 2nd International Conference on Knowledge Capture*, pages 70–77, Sanibel Island, Florida, USA.

Kamal Nigam and Matthew Hurst, 2005. *Computing Attitude and Affect in Text: Theory and Applications*, chapter Towards a Robust Metric of Opinion. Springer.

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

Jay M. Ponte and W. Bruce Croft. 1998. A language modeling approach to information retrieval. In *Proc. of the 21st Annual International ACM-SIGIR Conference*, pages 275–281, Melbourne, Australia.

James Shanahan, Yan Qu, and Janyce Wiebe, editors. 2005. *Computing attitude and affect in text*. Springer.

B. W. Silverman, 1986. *Density Estimation for Statistics and Data Analysis*, pages 75–94. CRC Press.

Ian Soboroff and Donna Harman. 2003. Overview of the TREC 2003 Novelty Track. In *Proc. of the 12th Text Retrieval Conference*, pages 38–53. NIST Special Publication 500-255.

Ian Soboroff. 2004. Overview of the TREC 2004 Novelty Track. In *Proc. of the 13th Text Retrieval Conference*. NIST Special Publication 500-261.

Fei Song and W. Bruce Croft. 1999. A general language model for information retrieval. In *Proc. of the 8th International Conference on Information and Knowledge Management*, pages 316–321, Kansas City, Missouri, USA.

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.

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. 2002. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 417–424, Philadelphia, Pennsylvania, USA.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 1(2):0–0.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proc. of HLT/EMNLP 2005*, Vancouver, Canada.

Jeonghee Yi, Tetsuya Nasukawa, Razvan Bunescu, and Wayne Niblack. 2003. Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. In *Proc. of the 3rd IEEE International Conference on Data Mining*, pages 427– 434, Melbourne, Florida, USA.

Chengxiang Zhai and John Lafferty. 2001. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proc. of the 24th Annual International ACM-SIGIR Conference*, pages 334–342, New Orleans, Louisiana, USA.

Chengxiang Zhai. 2002. *Risk Minimization and Language Modeling in Text Retrieval*. PhD dissertation, Carnegie Mellon University.