Opinionated Document Retrieval Using Subjective Triggers

Kazuhiro Seki

Organization of Advanced Science and Technology, Kobe University
1-1 Rokkodai, Nada, Kobe 657-8501, Japan
E-mail: seki@cs.kobe-u.ac.jp
Tel: +81-78-803-6480
Fax: +81-78-803-6316

Kuniaki Uehara

Graduate School of Engineering Kobe University
1-1 Rokkodai, Nada, Kobe 657-8501, Japan
E-mail: uehara@kobe-u.ac.jp
Tel: +81-78-803-6215
Fax: +81-78-803-6316

Please contact the first author, Kazuhiro Seki, for correspondence.

This paper proposes a novel application of a statistical language model to opinionated document retrieval targeting weblogs (blogs). In particular, we explore the use of the trigger model—originally developed for incorporating distant word dependencies—in order to model the characteristics of personal opinions that cannot be properly modeled by standard n-grams. Our primary assumption is that there are two constituents to form a subjective opinion. One is the subject of the opinion or the object that the opinion is about, and the other is a subjective expression; the former is regarded as a triggering word and the latter as a triggered word. We automatically identify those subjective trigger patterns to build a language model from a corpus of product customer reviews. Experimental results on the TREC Blog track test collections show that, when used for reranking initial search results, our proposed model significantly improves opinionated document retrieval. In addition, we report on an experiment on dynamic adaptation of the model to a given query, which is found effective for most of difficult queries categorized under politics and organizations. We also demonstrate that, without any modification to the proposed model itself, it can be effectively applied to polarized opinion retrieval.

Introduction

Since the advent of the Web, many forms of user-generated contents (UGC) have evolved, including personal homepages, discussion boards, and weblogs (blogs). Such UGC typically contains subjective opinions of individual authors which are difficult to find in the conventional mass media, such as magazines and newspapers. Among them, blogs have seen popularity as a means to express personal opinions regarding politics, hobbies, people, etc., due to the ease of use and maintenance. Because of its wide acceptance among the general public, blogs have been drawing much attention from natural language processing (NLP), information retrieval (IR), and other research communities as an attractive domain for exploration (Adar & Adamic, 2005; Agarwal et al., 2008; Ding et al., 2008; Esuli & Sebastiani, 2007; Mei et al., 2007).

Among a variety of research opportunities targeting blogs, this paper focuses on opinionated document (blog post) retrieval, a task to retrieve blog posts not only relevant to a user query but also containing subjective opinions of authors. Opinionated document retrieval has been challenged by many researchers, partly motivated by the Text Retrieval Conference (TREC) Blog track (Macdonald et al., 2007; Ounis et al., 2006, 2008). Previous works by the track participants and others can be roughly categorized into lexicon-based (Lee et al., 2008; Mishne, 2006; Oard et al., 2006; Vechtomova, 2010; Zhang & Ye, 2008) and classification-based (Gerani et al., 2009; Zhang & Yu, 2006; Zhang et al., 2007). Briefly, the former uses a manually or automatically compiled list of words, such as "like" and "fantastic", and in essence assumes the existence of those words in a document as an indicator of opinions. The latter, classification-based, also typically relies on word occurrences but automatically create a classifier based on positive (i.e., opinionated) and negative (i.e., non-opinionated) examples using machine learning algorithms.

In this paper, we propose a simple but effective approach to opinionated document retrieval (or opinion retrieval for short) which does not belong to either category. Our approach was in part inspired by the empirical finding that considering the proximity of pronouns and subjective expressions to objects improves opinion retrieval (Zhou et al., 2007). We take advantage of statistical language models for capturing such characteristic patterns often seen in opinionated documents. In particular, we explore the use of the classic trigger model (Lau et al., 1993; Tillmann & Ney, 1996), which was originally proposed for dealing with long-distance word dependencies. Our primary assumption is that there are two essential constituents to form a personal or subjective opinion. One is the subject of the opinion (e.g., "I") or the object that the opinion is about (e.g., "this movie"), and the other is a subjective expression (e.g, "like"). We regard the former as a triggering word and the latter as a triggered word and automatically identify trigger patterns characteristic to subjective opinions using customer reviews collected from `Amazon.com`. Through several experiments on the Text Retrieval Conference (TREC) Blog track test collections, it is demonstrated that, when used for reranking, our proposed model significantly improves IR system performance and that dynamically adapting the model to a given query gives steady improvement. Also, it is shown that our approach can be easily extended to polarized document retrieval which distinguishes positive and negative opinions.

In the rest of this paper, we first detail our approach to building a trigger model for subjective opinions. Then, we evaluate the validity of our proposed model and its effectiveness in retrieving opinionated blog posts by way of a variety of experiments on the Blog track test collections. After that, we summarize the related work. Lastly, we

conclude this paper with a brief summary of the findings and possible future directions.

Opinion Retrieval by a Trigger Model

Motivation

To judge whether a given document contains subjective opinions, the simplest and most intuitive approach would be to look for subjective words in the document. The underlying assumption of this kind of lexicon-based approaches is that if a document contains words often used for expressing subjectivity, it is likely to be opinionated. For instance, "like" may be a good indicator for favorable feelings. Along this line, many researchers manually or automatically created a sentiment-oriented word list or dictionary to use for identifying opinions (for example, Lee et al., 2008; Vechtomova, 2010; Yang et al., 2006). Although reported effective, a potential limitation of this approach is, as opposed to the intuition, that a document with such subjective words is not necessarily opinionated. For example, "It looks *like* a cat." or "She *likes* singing." may possibly be a subjective opinion of the writer but is rather objective; the former uses "like" as a preposition and the latter is a statement or a fact about a third person. To distinguish such difference, one would need to look at wider context wherein those potentially subjective words occur.

One way to consider wider context is to use the classic n-gram language models (Manning & Schütze, 1999), which estimate the probability of a word occurrence based on the prior local context. Basically, it treats n consecutive terms as a unit of analysis. For example, bigrams in the above sentence "It looks like a cat." are "It looks", "looks like", "like a", and "a cat", where "like" is now analyzed with the local context (i.e., "looks" and "a"), rather than the individual occurrence. Although one could take into account as wide context as she wants, simply increasing n will cause data sparseness and result in unreliable parameter estimation. For such reasons, n is often set to 2 or 3 depending on the intended application and the amount of available corpora.

In this work, we aim to improve opinionated document retrieval and study the use of trigger models for capturing patterns or word dependencies that are characteristic to subjective opinions.

Subjective Trigger Models

Despite its simplicity, n-gram language models have been successfully applied to many NLP-related problems. However, it is clear that there exist long-distance dependencies beyond the limited horizon specified by n. To include such dependencies, Lau et al. (1993) proposed the trigger-based language model (or trigger model for short). A trigger refers to a word that tends to bring about the occurrence of the other. For example, "neither" and "nor" are often used as a pair in the same sentence, such as "I am *neither* a liberal *nor* a conservative". (We will use this example in the following to illustrate the trigger model.) An n-gram model is not suitable for capturing this kind of dependencies because the words between "neither" and "nor" can be any phrases with any length. A trigger model $P_T(w|h)$ could incorporate such trigger pairs and is used to enhance a baseline n-gram language model $P_B(w|h)$ by linearly interpolating the two:

$$P_E(w|h) = (1 - \lambda) \cdot P_B(w|h) + \lambda \cdot P_T(w|h) \quad (1)$$

where w and h denote a word and a history, respectively, and λ is the interpolation parameter. For the above example of "nor", the baseline n-gram model P_B's history h is its preceding words (e.g., "a liberal" for $n = 3$), and for the trigger model P_T's history may be "neither". We will briefly describe the definition of $P_T(w|h)$ later.

To build a trigger model, we first need to identify significant triggering and triggered word pairs (e.g., "neither" and "nor"). Given a corpus of documents, any word pair, such as "I → nor" and "am → nor", in the vocabulary can potentially be a trigger pair. Here, vocabulary is a list of words that appear in a given corpus. Tillmann & Ney (1996) proposed a criterion to consider word w as a potential triggered word only when an n-gram model $P(w|h)$ without smoothing (different from $P_B(w|h)$) gives "poor" estimation for w, meaning that $P(w|h)$ is smaller than a predefined threshold t. That is,

$$P(w|h) < t \,. \tag{2}$$

For example, if P(nor|a liberal) is smaller than t, "nor" is considered as a potential triggered word b. In other words, the exact word sequence "a liberal nor" rarely appears in a given corpus, we look for a better triggering word that can predict an occurrence of "nor", such as "neither", appearing out of the window of n-grams. How far we look for a triggering word (i.e., the size of history h), is arbitrarily chosen and is typically limited to a sentence, a paragraph, or a document at most.

Each word b satisfying Equation (2) is evaluated in combination with every word a in the vocabulary to determine whether any pair "$a \rightarrow b$" provides better estimation based on the log-likelihood difference between an n-gram language model $P(\cdot)$ and a mixture model enhanced *only* by the pair "$a \rightarrow b$" under consideration, denoted as $P_{E:a\rightarrow b}(\cdot)$. More precisely, given input texts (the entire corpus) represented as a long word sequence $w_1, w_2, \ldots, w_m$, the difference $\Delta_{a\rightarrow b}$ is computed as follows.

$$\begin{aligned} \Delta_{a\rightarrow b} &= \log P_{E:a\rightarrow b}(w_1, w_2, \ldots, w_m) - \log P(w_1, w_2, \ldots, w_m) \\ &\approx \sum_i \log \left(P_{E:a\rightarrow b}(w_i|h_i) - P(w_i|h_i)\right) \end{aligned} \tag{3}$$

For the neither-nor example, the pair is evaluated by the following.

$$\Delta_{\text{neither}\rightarrow\text{nor}} \approx \sum_i \log(P_{E:\text{neither}\rightarrow\text{nor}}(w_i|h_i) - P(w_i|h_i)) \tag{4}$$

Note that the extended language model $P_{E:\text{neither}\rightarrow\text{nor}}(\cdot)$ is composed of the baseline language model $P_B(\cdot)$ and the trigger model $P_T(\cdot)$ as in Equation (1), but here $P_T(\cdot)$ is estimated by looking at only "neither" as history h, so as to evaluate the utility of "neither" to predict the occurrence of "nor". Simply put, if "neither" often appears with "nor" in the given corpus as compared with immediately preceding $(n-1)$ words, $\Delta_{\text{neither}\rightarrow\text{nor}}$ becomes greater. After evaluating each possible word pair like neither→nor, one can take an arbitrary number of pairs with greatest log-likelihood difference to build the final trigger model $P_T(\cdot)$. This criterion, or the triggers identified by the criterion, is called the *low level triggers*. Using this criterion, Tillmann & Ney identified trigger pairs, such as "neither → nor" and "tip → iceberg".

We adopt the trigger model with some modification described shortly for capturing the characteristics of subjective opinions based on two assumptions. The first, primary assumption is that a subjective opinion usually contains

two essential components: the subject of the opinion (e.g., "I") or the object that the opinion is about (e.g., "this movie") and a subjective expression (e.g., "like" and "best"). We regard the former as the triggering word and the latter as the triggered word. The second assumption is that the triggering word often appears as a pronoun. These assumptions reflect the empirical finding that proximity of pronouns (e.g., "I", "you", and "me") and subjective expressions (e.g., "like" and "feel") to objects is an effective measure of opinionatedness (Yang et al., 2007; Zhou et al., 2007). By introducing these assumptions, we could acquire trigger pairs, such as "I → really" and "I → like", which may be useful to identify opinionated documents.

In contrast to the ad hoc heuristics used in the previous work, our model provides a more principled way to incorporate the term dependencies indicating opinionatedness. Also, by only considering a set of pronouns as potential triggering words, we can build both more efficiently and more effectively a focused language model tailored to personal subjective opinions. In the following, we call the language model enhanced by the subjective triggers the *subjective trigger model*.

Building a Subjective Trigger Model

Based on the procedure and assumptions described in the previous section, we preliminarily built a subjective trigger model as follows. First, we identified trigger pairs potentially representing subjective opinions. For this purpose, we needed a corpus consisting of subjective opinions. This study used 5,000 customer reviews automatically collected from `Amazon.com`. The corpus size is 4.1 MB, containing 785,626 word tokens in total and 25,292 unique word types.[1] These reviews were written for various kinds of products sold at Amazon, including books, DVDs, electrical appliances, toys, etc. Here, their customer ratings (ranging from 1 to 5) were not distinguished because they are all supposed to be subjective opinions whether positive, negative, or neutral.

As potential triggering words, we experimentally chose 14 pronouns: `I`, `my`, `you`, `it`, `its`, `he`, `his`, `she`, `her`, `we`, `our`, `they`, `their`, and `this`, and identified 2,138 trigger pairs using the low level triggers criterion with the threshold t set to e^{-6}.[2] In building the model, we limited the history h to the prior context (preceding words) in the same sentence. Table 1 shows the trigger pairs with highest log-likelihood improvement (see Equation (3)).

As shown in Table 1, the trigger pairs identified by the low level trigger criterion are dominated by function words, seemingly not very useful for characterizing opinionated documents. This problem is caused by the fact that function words appear in many context, which sometimes leads to a low probability $P(w|h)$. More precisely, function words, such as "the", occur after many different histories, and Equation (2) is evaluated for each unique history h. For a given particular corpus, there would be some h's after which "the" occurs rarely, resulting in lower $P(\text{the}|h)$ than threshold t. Another problem regarding the criterion is that it does not directly evaluate the frequency of history, $Freq(h)$. In general, $Freq(h)$ needs to be sufficiently high in order to obtain reasonable estimate for $P(w|h)$. To incorporate these two factors, we modified the criterion as follows:

$$\tau \cdot P(w_i|h_i) < t \tag{5}$$

[1] We also tested larger corpora but it made no significant difference in the trigger pairs identified or the performance of opinion retrieval.

[2] Varying t did not generate better subjective trigger pairs, although it gave slightly different ones.

Table 1: Most prominent low level triggers

Triggering (a)		Triggered (b)	$\Delta_{a\rightarrow b}$
`this`	→	`the`	7.079
`it`	→	`the`	7.079
`i`	→	`the`	7.079
`i`	→	`to`	6.526
`this`	→	`to`	6.525
`my`	→	`and`	6.502
`i`	→	`and`	6.501
`this`	→	`and`	6.498
`it`	→	`and`	6.497
`this`	→	`a`	6.381
		...	

where τ is defined as the ratio of the frequency of w_i to that of h_i, i.e., $Freq(w_i)/Freq(h_i)$. This modification penalizes frequent words w_i with infrequent history h_i to prevent (mainly) function words from being identified as triggered words. Alternatively, one could use a precompiled stopword list, which, however, may result in missing useful trigger pairs involving function words, such as "this → the" as in "This is the best choice." and "this → all" as in "This book is all you need." The former characterizes the superlative degree and the latter is an idiomatic expression for recommendation. Note that blogs are often written informally and "the" may be dropped (e.g., "this is best"). Also, "best" on its own would be a good opinion cue which cannot be modeld by a trigger model.[3] Furthermore, "this → all" may pick up non-opinion expressions, such as "this is all I have". Despite these potential limitations, the *Evaluation* section will later demonstrate that our proposed model is effective for retrieving opinionated blogs.

Table 2 shows the most prominent trigger pairs with highest log-likelihood improvement among 2,012 pairs identified by the modified criterion with the thereshold t set to e^{-8}. We can observe that many of the trigger pairs appear to be characteristic to personal opinions and that some pairs were still able to involve function words. Although not shown here, we can find other trigger pairs further down the list, which capture more distant word dependency, including "`I` → `very`" and "`it` → `greatest`". In order to verify that our approach indeed captures distant word dependencies, Figure 1 shows a histogram of distance (measured as the number of words) between the identified triggering and triggered word pairs. In addition, Figure 2 shows the same only for shorter distance. The results indicate that many of the trigger pairs could not be found by an n-gram model with a small value of n, such as 3.

An alternative approach to identifying a similar set of word pairs would be to employ a syntactic parser. Our proposed framework has two advantages over using a parser. First, because our framework does not require NLP tools and relies only on word occurrences, it is more easily applicable to other languages as long as the same assumptions apply; that is, a subjective opinion has two constituents, one of which is a pronoun. Second, since it

[3] Although a trigger model does not consider individual terms, the baseline language model combined with a trigger model is a back-off n-gram model, which uses lower-order n-grams including individual terms (unigrams) when needed.

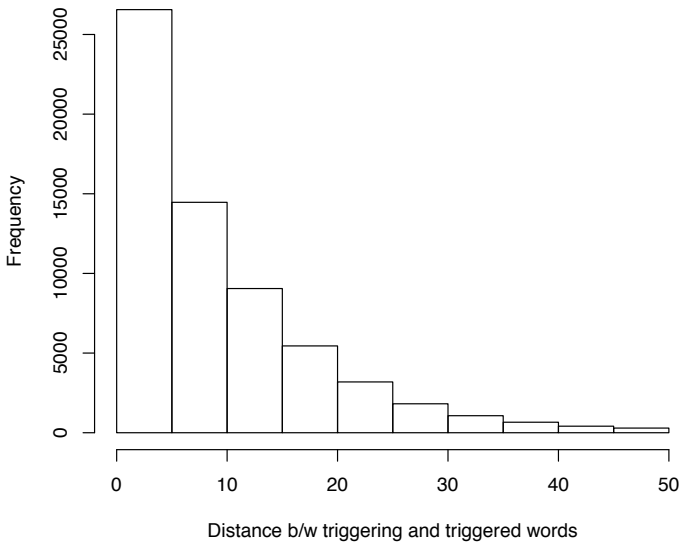

Figure 1. Histograms of the distances (number of words) between identified triggering and triggered words.

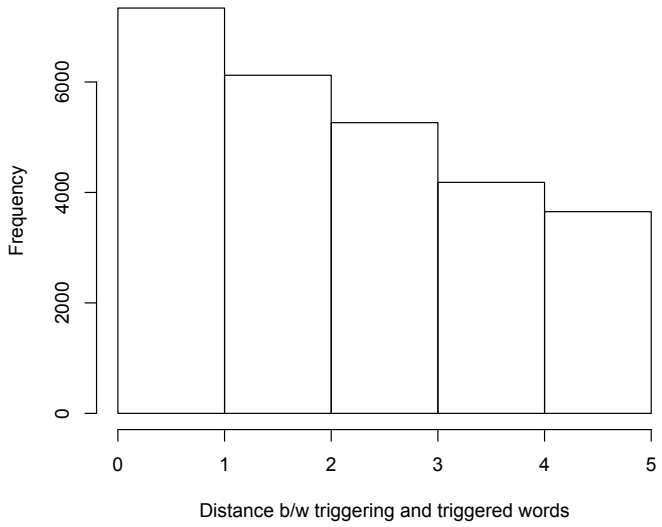

Figure 2. Histograms of the distances (1 to 5) between identified triggering and triggered words.

Table 2: Most prominent triggers identified by the modified criterion

Triggering (a)		Triggered (b)	$\Delta_{a \to b}$
`i`	→	`wish`	5.113
`i`	→	`felt`	5.073
`i`	→	`loved`	4.862
`i`	→	`hope`	4.739
`i`	→	`couldn`	4.680
`i`	→	`got`	4.611
`i`	→	`cannot`	4.593
`this`	→	`an`	4.578
`this`	→	`all`	4.575
`i`	→	`liked`	4.552
`i`	→	`enjoyed`	4.531
		...	

does not consider dependency relations, it is capable of discovering not directly dependent word pairs. For example, the trigger pairs identified above include "`I` → `fantastic`" as in, e.g., "I thought the writing is fantastic".

After identifying trigger pairs, the association score between a and b, $\alpha(b|a)$, for each identified trigger pair ($a \to b$) is calculated as the maximum likelihood estimate based on their collocations as follows:

$$\alpha(b|a) = \frac{N(a,b)}{N(a)} \tag{6}$$

where $N(a,b)$ is the number of occurrences of b for which word a appears in b's history and $N(a)$ is the number of occurrences of a in any histories. Then, the trigger model $P_T(w|h)$ is built as follows:

$$P_T(w|h) = \frac{1}{|h|} \sum_{w_j \in h} \alpha(w|w_j). \tag{7}$$

Equation (7) basically states that $P_T(w|h)$ is estimated by averaging the association scores, $\alpha(w|w_j)$, for every combination of w and w_j, where w_j is a history word of w. For more details of the estimation method, readers are referred to Tillmann & Ney (1996)'s paper. As the baseline language model, we used a smoothed, back-off trigram model, $P_B(w|h)$, and empirically set $\lambda = 0.9$ in Equation (1), giving a higher weight to the trigger model.

We do not claim that our proposed model can capture all the clues that indicate subjective opinions. For example, subjective expressions may be malformed (e.g., "fantaaaaaastic"), which are unlikely to exist in the vocabulary of our language model. Also, subjective trigger pairs may be about something other than the target of interest. For example, suppose a query "Bush" returned a document containing the next sentence: "Politics, I hate it, but here are the facts on Bush." It has a typical trigger pair "I → hate", which is, however, not directly talking about Bush. Dealing with these cases would require other techniques, such as lexicon- and NLP-based, and is left for possible future work.

Model Adaptation

Since the subjective trigger model was built on Amazon customer reviews, which essentially deal only with products, it may not be very effective to identify opinionated documents on some types of topics or queries other than products. To tackle the potential drawback, we propose the adaptation of the trigger model by identifying additional trigger pairs in the blog posts returned by initial search.

The idea is in essence similar to pseudo-relevance feedback (PRF) (Lavrenko & Croft, 2001; Sakai et al., 2005; Tao & Zhai, 2006) which expands the original query by adding useful terms found in top k blog posts in the initially retrieved set. An important difference between PRF and our adaptation method is that we do *not* modify the original query but updates the language model for better estimating the opinionatedness of a given blog post. Thus, PRF can also be applied irrespective of the model adaptation if desired.

The following describes the procedure of our proposed model adaptation.

1. Carry out initial search by a choice of an IR model for a given user query q (e.g., "Skype"). Note that the IR model used in this step can be any general IR model which returns an ordered list of documents for the given q.

2. Using the top k blog posts retrieved, identify trigger pairs, $a \rightarrow b$, and compute their associations, $\alpha'(b|a)$ in the same way described above. It should be emphasized that the k blog posts are used instead of Amazon customer reviews to dynamically learn triggering-triggered word associations specifically for query q. Because this step is query-specific, one could use q itself (e.g., "Skype") as potential triggering words in addition to the predefined set of 14 pronouns (see the *Building a Subjective Trigger Model* section) so as to capture sentences containing q itself as the subject as in "Skype is the best". We will later show how the choice of triggering words influences the opinion retrieval performance.

3. Estimate the trigger model $P_T(\cdot)$ using either $\alpha(b|a)$ (the original term associations learned offline) or $\alpha'(b|a)$ (learned in the previous step) with a greater value. That is, instead of Equation (7), we use Equation (8).

$$P_T(w|h) = \frac{1}{|h|} \sum_{w_j \in h} \max\left(\alpha(w|w_j), \alpha'(w|w_j)\right) \tag{8}$$

Intuitively, if strong associations $\alpha'(w|w_j)$ between triggering and triggered words are found in the top k blog posts returned for query q, they overwrite default associations $\alpha(w|w_j)$ learned from a larger corpus of opinions (Amazon customer reviews) which may or may not be appropriate for the particular query q.

This adaptation enables to incorporate prominent trigger pairs based on the top k blog posts into the subjective trigger model. Although topically relevant documents are not necessarily opinionated and thus using top k blog posts may not be well justified, blogs are often subjective by nature. In fact, strong correlation between the performance of initial search and opinion retrieval has been repeatedly reported in the literature (Macdonald et al., 2007; Ounis et al., 2006, 2008).

Topic #	851
Title	March of the Penguins
Desc.	Provide opinion of the film documentary “March of the Penguins”.
Narr.	Relevant documents should include opinions concerning the film documentary “March of the Penguins”. Articles or comments about penguins outside the context of this film documentary are not relevant.

Figure 3. Example topic from the TREC Blog track.

Evaluation

Data Set

To evaluate the validity of the proposed model, we conducted evaluative experiments on the Blog06 corpus created for the TREC Blog track (Ounis et al., 2006). It is a collection of over 3.2 million blog posts crawled over an 11 week period from December 2005 to February 2006. As user information needs, we used 150 topics provided for the Blog track 2006 to 2008 (50 topics per year). The topic sets for 2006 and 2007 were developed from commercial blog search engine query logs, and those for 2008 were developed by assessors at the National Institute of Standards and Technology. Figure 3 gives an example topic, where “topic #” indicates the identification number of the topic, “title” is (or resembles) the actual query used by search engine users, “desc.” is the description of the topic, and “Narr.” is the more detailed account of the information needs.

For each topic, relevant/irrelevant blog posts in the Blog06 corpus are marked (not exhaustively) by the pooling method (Voorhees & Harman, 2005) for evaluating a given IR system. Relevance judgment has been done in five categories: irrelevant (labeled as 0), relevant and not opinionated (1), relevant and only negatively opinionated (2), relevant and both positively and negatively (or neutrally) opinionated or (3), and relevant and only positively opinionated (4). Note that, for a standard IR evaluation, labels 1–4 are not distinguished and treated as a single “relevant” category, whereas we consider only the labels 2–4 as relevant in the context of opinion retrieval.

Using the Blog track test collection, we evaluated our proposed model in two ways as detailed in the following sections. First, we assessed the validity of the language model itself out of the context of IR. Second, we examined the effectiveness of the model for opinion retrieval in an IR setting, followed by an evaluation of the model adaptation.

Evaluation of the Language Model

The subjective trigger model integrating 2,012 trigger pairs built in the *Building a Subjective Trigger Model* section was based not on opinionated blogs but on Amazon customer reviews, which may not be ideal resources as they are reported to contain many “spam” reviews (Jindal & Liu, 2008). Therefore, we first examined whether the subjective trigger model was able to reflect the characteristics of opinionated blog posts. For this purpose, we used a measure called *perplexity* commonly used for evaluating language models (Jurafsky & Martin, 2000). Intuitively, perplexity

Table 3: Perplexity results. Figures in parentheses indicate percent decrease of perplexity as compared to corresponding P_B

n	Non-opinionated (d_N)		Opinionated (d_O)	
	P_B	P_E	P_B	P_E
1gram	9369	8946 (−4.5%)	7198	6829 (−5.1%)
2gram	6526	6279 (−3.8%)	4749	4546 (−4.3%)
3gram	5998	5762 (−3.9%)	4337	4145 (−4.4%)

quantifies how much uncertainty a language model leaves in predicting a word sequence (document), and thus, lower perplexity generally means a better model. More formally, perplexity is defined as $2^{H(P,d)}$, where $H(P, d)$ denotes cross entropy of language model P on document $d = w_1\ w_2\ \ldots w_m$ (a sequence of words). Cross entropy is an information theoretic measure of distance between an estimated and true probability distributions and defined as in Equation (9) when m approaches to ∞.

$$H(P, d) = -\frac{1}{m} \log P(w_1 \ldots w_m) \tag{9}$$

When the language model P is approximated by an n-gram model (or a trigger model) in which the probability of word (w_i) occurrence is conditioned on its history h_i (e.g., $n - 1$ preceding words for n-grams), cross entropy is estimated as in Equation (10).

$$H(P, d) \approx -\frac{1}{m} \sum_{i=1}^{m} \log P(w_i|h_i) \tag{10}$$

Using Equation (10), we can compare the perplexities of different language models on the same document d and determine which model is more accurate.

We concatenated all the opinionated blog posts labeled 2–4 (i.e., relevant and opinionated) to create a single very long document d_O, and similarly created another document d_N from all the non-opinionated blog posts labeled 1 (i.e., relevant only). For this experiment, we looked at the blog posts associated with only the 50 topics for the 2006 Blog track. Table 3 presents perplexity results for the baseline language model P_B and the subjective trigger model P_E with different n.

In the results, we can make three important observations. First, with higher order n-grams, perplexity monotonically decreases irrespective of language models and document types, which means that a language model with higher n, at least up to 3, better represents opinionated documents. Second, opinionated document d_O lead to lower perplexity than non-opinionated document d_N. This result suggests that the language models learned from Amazon customer reviews capture some characteristics of opinions in blogs. Third, the subjective trigger models P_E produce lower perplexity than the baseline language models P_B and the difference is slightly greater for opinionated document d_O. It may indicate that the subjective trigger pairs brought some additional clues of opinionatedness that could not be captured by standard n-gram models.

The above experiment implies the potential effectiveness of the subjective trigger models for discriminating opinionated documents from the non-opinionated ones at large. Then, we investigated if it holds at the individual

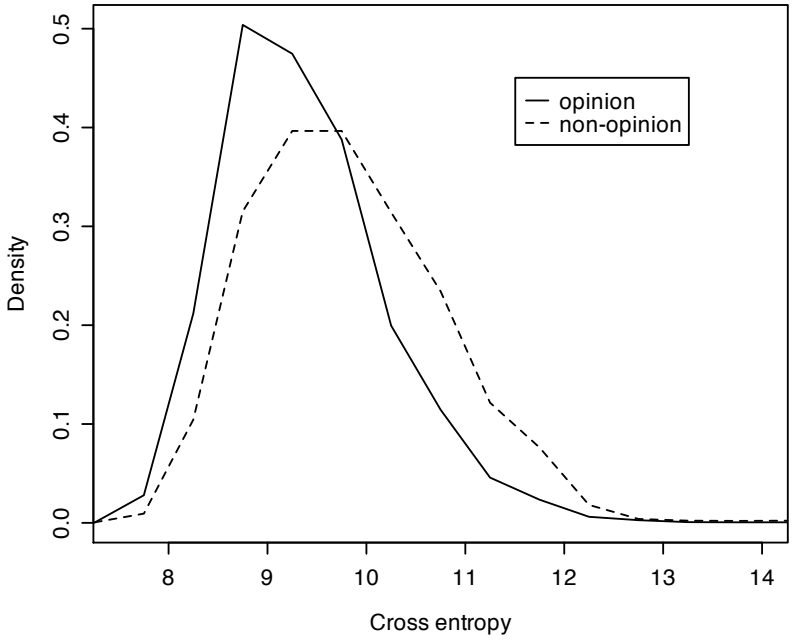

Figure 4. Distributions of cross entropy for opinionated and non-opinionated blog posts.

document level by comparing the distributions of cross entropy of P_E on opinionated and non-opinionated blog posts, where n was set to 3. Figure 4 shows the result, where x axis is cross entropy and y axis is probability density such that the area under the curve equals one. The result confirms that, using the proposed model, the distribution of cross entropy for opinionated blog posts generally takes lower values than that for non-opinionated ones. The difference between their means is statistically significant by two-sided t-test at the 0.01 level ($p = 2.26 \times 10^{-16}$).

Next section reports on another set of experiments on opinion retrieval by integrating the subjective trigger model into a general IR system through document reranking.

Opinion Retrieval with the Subjective Trigger Model

Initial Retrieval

We evaluated the effectiveness of our proposed model for opinion retrieval by applying it to initial search results returned by a general IR model. For initial search, we tested two alternatives: a) the vector space model (Salton & McGill, 1983) with the TFIDF term weighting (Sparck Jones, 1972), denoted as VSM, and b) the inference network model combined with a language modeling approach (Metzler & Croft, 2004), denoted as LM. For both models, indexing was done case insensitively after removing stopwords, where no stemmer was applied. For queries, we used only topic titles (see Figure 3). Remember that we consider the labels 2–4 as relevant disregarding the polarity of opinions, i.e., positive, negative, and mixed (see the *Data Set* section). Table 4 shows the initial retrieval results in mean average precision (MAP) for the three topic sets (2006 to 2008) and also shows those produced by Lee et

Table 4: Initial search results using alternative IR models, where the official results for the best baseline run, baseline4, are shown for reference

Model	MAP			
	2006	2007	2008	All
VSM	0.1126	0.1671	0.1727	0.1508
LM	0.1965	0.2458	0.2475	0.2299
baseline4 (Lee et al., 2008)	0.3022	0.3784	0.3822	0.3543

al. (2008), referred to as "baseline4", which marked the best performance at the 2008 Blog track.[4] Briefly, baseline4 was also produced by a similar language modeling approach as the LM above but achieved greater performance by considering not only document-level relevance but also passage-level relevance. In addition, Lee et al. extracted the body text of each blog post by looking at the difference between blog posts within the same blog site, whereas we extracted all the texts appearing in blog posts.

There is a large difference in performance between the two alternative IR models we tested, i.e., VSM and LM. Due to the observed disadvantage, the following experiments do not use VSM and attempt to refine the initial search results of LM and baseline4 in terms of opinion retrieval.[5]

Integration of a Subjective Trigger Model

In the initial retrieval by LM or baseline4, each retrieved blog post d is assigned a probability $P(I|d)$ that a given d is relevant to user's information need I. Assuming that whether a given d is opinionated is independent of being topically relevant to I, the probability that d is both topically relevant and opinionated can be expressed as a product of $P(I|d)$ and $P_E(d) \approx \prod_i P_E(w_i|h_i)$. However, because longer blog posts tend to have smaller $P_E(d)$ by definition and the two probability distributions may have largely different variances, simply multiplying the two generally does not work. Thus, we take the weighted sum of their logarithms and normalize $P_E(d)$ by the document length (word count) m to produce the final score, $Scr(d, I)$, to rerank the blog posts:

$$Scr(d, I) = (1 - \beta) \cdot \log P(I|d) + \frac{\beta}{m} \sum_i \log P_E(w_i|h_i) \tag{11}$$

where β is an interpolation parameter controlling the effect of the language model enhanced by subjective triggers. Notice that the second term corresponds to cross entropy in Equation (9). For IR models which do not provide probabilities (e.g., VSM), an alternative score can be defined as some form of linear combination or by more theoretical fusion techniques (Zhang & Ye, 2008).

[4]Five baseline runs including baseline4 were made available at `http://trec.nist.gov/data/blog08.html` so as to enable direct comparison across different opinionated document ranking algorithms.

[5]It does not mean that LM is superior to VSM. In fact, Zhang et al. (2007) obtained much higher MAP than our LM by employing a vector space model.

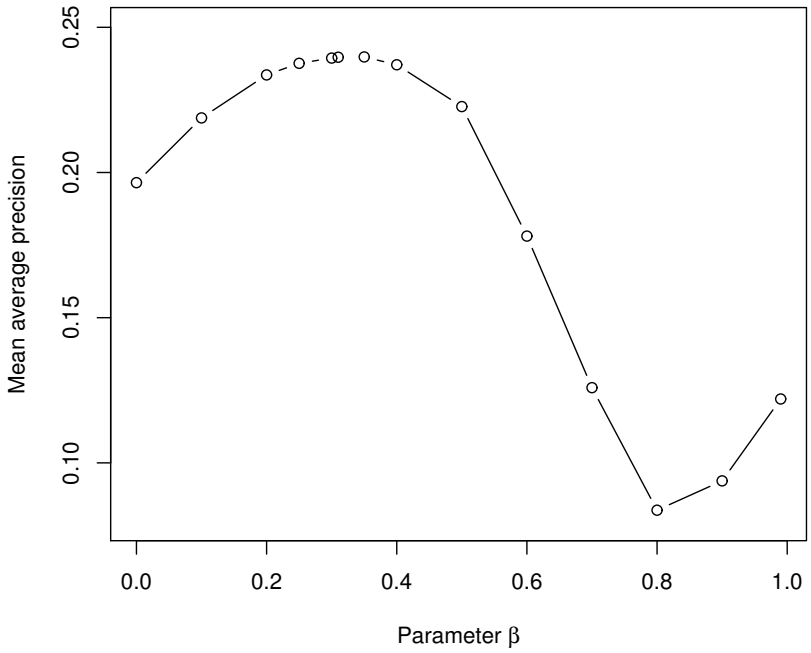

Figure 5. Relation between parameter β and MAP.

Table 5: Performance of opinion retrieval in MAP before/after reranking by subjective trigger model

Model		MAP			
		2006	2007	2008	All
baseline4	Before	0.3022	0.3784	0.3822	0.3543
	After	0.3259 (+7.84%)	0.4078 (+7.77%)	0.4036 (+5.60%)	0.3791 (+7.00%)
LM	Before	0.1965	0.2458	0.2475	0.2299
	After	0.2398 (+22.0%)	0.3280 (+33.4%)	0.3057 (+23.5%)	0.2920 (+27.0%)

We gradually increased the parameter β from 0 to 1 in Equation (11) and reranked the documents initially retrieved by LM to see if any improvement in MAP is observed. Figure 5 shows the transition of the MAP score for different values of β for the 50 topics from the 2006 Blog track. The leftmost circle, where $\beta = 0$, corresponds to the initial result. By varying β, MAP increased by 0.2398 (+22.0%) as compared to the initial result (MAP = 0.1965). This observation verifies that the subjective trigger model integrated through Equation (11) is effective for spotting opinionated blog posts without degrading the initial topic-based ranking if β is properly chosen. Table 5 summarizes the performance in MAP after reranking for the 2006 to 2008 topic sets. Even with the strongest baseline4 from the Blog track, the reranked results improved MAP by 7.00% overall. It should be also emphasized that the interpolation parameter β was found optimal at around 0.35 for LM across all the data sets, which implies desirable stability of the optimum value of β across different topics. Similarly, the optimum β for baseline4 was stable across all the data sets.

Analysis on Individual Queries

The previous section reported that incorporating the trigger model into a reranking scheme improved MAP by 7.00% and 27.0% on average when baseline4 and LM was used for initial retrieval, respectively. While the overall effect was positive, it does not tell how the trigger model affected the result for each topic. To shed light on it, we looked at individual topics. Figure 6 summarizes the average precision improvement (difference) as a bar plot, where each bar corresponds to a topic and a positive value indicates an improvement after reranking by our trigger model. The topics are in ascending order of their topic numbers from left to right. For all the topics, the parameter β was fixed to the optimum identified above. We can see that our approach was effective for the majority of the topics. For baseline4 (the strongest baseline), 81 cases out of 150 achieved 5% or more improvement, whereas 13 cases had 5% or more performance reduction. For our own baseline, LM, the numbers are 120 and 21, respectively.

Then, we further analyzed the results, focusing on the 50 topics from the 2006 Blog track. Table 6 shows the details when baseline4 was used for initial retrieval, where "Initial" and "Trigger" are the average precision for each topic before and after reranking, respectively, and "Difference" and "% imprv" are the improvement from "Initial" shown as difference and percentage, respectively.

Examining the results, notable increase (> 0.05 points in AP) was observed for **macbook pro** (+0.0634), **mardi gras** (+0.0811), **natalie portman** (+0.0678), **shimano** (+0.2622), **zyrtec** (+0.1057), and **board chess** (+0.0647). On the other hand, there is a slight performance drop (> 0.001 points in AP) for the following topics: **ann coulter** (−0.0267), **letting india into the club?** (−0.0142), **basque** (−0.0020), **barry bonds** (−0.0094), **brokeback mountain** (−0.0136), **sonic food industry** (−0.0026), **fox news report** (−0.0067), **seahawks** (−0.0250), **world trade organization** (−0.0086), and **jim moran** (−0.0016). To help identify the commonalities, if any, underlying respective topic sets, the following summarizes their descriptions, mostly from Wikipedia.[6]. Regarding the topics for which notable improvement was observed:

- **MacBook Pro** (#856) is a line of Macintosh portable computers by Apple Inc. for the professional, gaming and power user market.
- **Shimano** (#885) is a Japanese multinational manufacturer of cycling components, fishing tackle, and snowboarding equipment.
- **Zyrtec** (#893) is a medication that is used to treat allergy symptoms and chronic hives.
- **Mardi Gras** (#861) is the final day of Carnival, the three day period preceding the beginning of Lent, the Sunday, Monday, and Tuesday immediately before Ash Wednesday.
- **Natalie Portman** (#880) is an Israeli-American actress.
- **Board chess** (#894) is the traditional game of chess using 32 pieces and played on a board having 64 black and white squares.

The first half of the above topics, MacBook Pro, Shimano, and Zyrtec, can be categorized as products. Note that Shimano is a company name but is also often used to refer to their products. Even though the Amazon customer reviews used to build the trigger model do not specifically talk about these topics, such as medication, the model

[6] `http://en.wikipedia.org/wiki/Wiki`

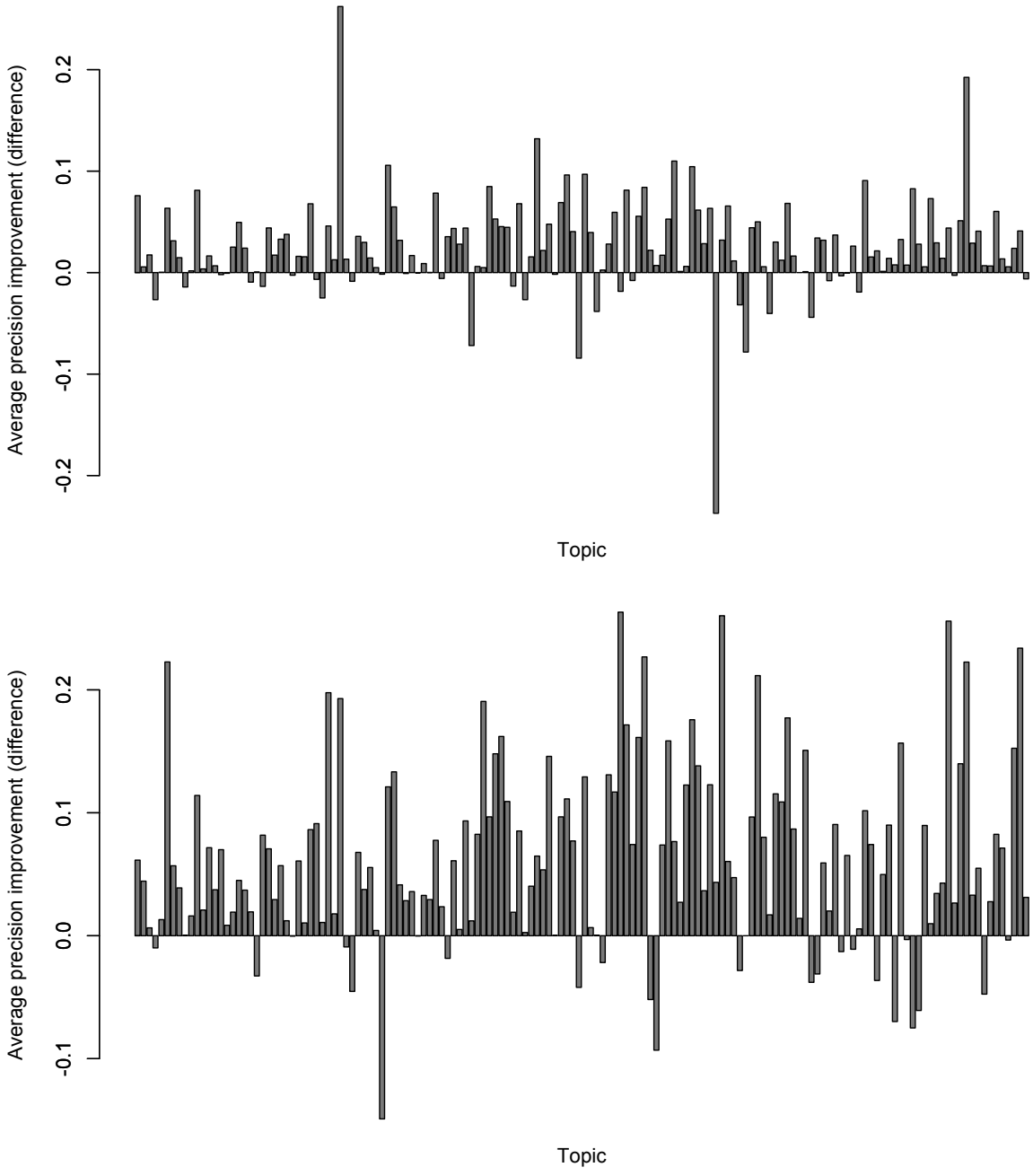

Figure 6. Average precision improvement over baseline4 (top) and over LM (bottom) after reranking for individual topics from the 2006 to 2008 Blog track.

Table 6: Individual results for the topics from the 2006 Blog track in average precision comparing initial retrieval (by baseline4) and one after reranking.

Topic #	Topic	Initial	Trigger	Difference	% imprv
851	march of the penguins	0.4348	0.5106	+0.0758	17.4%
852	larry summers	0.5469	0.5525	+0.0056	1.0%
853	state of the union	0.1633	0.1807	+0.0174	10.7%
854	ann coulter	0.5715	0.5448	−0.0267	−4.7%
855	abramoff bush	0.4735	0.4738	+0.0003	0.1%
856	macbook pro	0.4412	0.5046	+0.0634	14.4%
857	jon stewart	0.3605	0.3918	+0.0313	8.7%
858	super bowl ads	0.2083	0.2230	+0.0147	7.1%
859	letting india into the club?	0.5799	0.5657	−0.0142	−2.4%
860	arrested development	0.3354	0.3371	+0.0017	0.5%
861	mardi gras	0.3299	0.4110	+0.0811	24.6%
862	blackberry	0.0502	0.0538	+0.0036	7.2%
863	netflix	0.5967	0.6131	+0.0164	2.7%
864	colbert report	0.4488	0.4555	+0.0067	1.5%
865	basque	0.4285	0.4265	−0.0020	−0.5%
866	whole foods	0.0348	0.0341	−0.0007	−2.0%
867	cheney hunting	0.3179	0.3430	+0.0251	7.9%
868	joint strike fighter	0.1611	0.2106	+0.0495	30.7%
869	muhammad cartoon	0.2768	0.3009	+0.0241	8.7%
870	barry bonds	0.3347	0.3253	−0.0094	−2.8%
871	cindy sheehan	0.3956	0.3962	+0.0006	0.2%
872	brokeback mountain	0.3861	0.3725	−0.0136	−3.5%
873	bruce bartlett	0.4132	0.4572	+0.0440	10.6%
874	coretta scott king	0.5049	0.5221	+0.0172	3.4%
875	american idol	0.2498	0.2827	+0.0329	13.2%
876	life on mars	0.1772	0.2149	+0.0377	21.3%
877	sonic food industry	0.0398	0.0372	−0.0026	−6.5%
878	jihad	0.1144	0.1304	+0.0160	14.0%
879	hybrid car	0.1784	0.1940	+0.0156	8.7%
880	natalie portman	0.4349	0.5027	+0.0678	15.6%
881	fox news report	0.4256	0.4189	−0.0067	−1.6%
882	seahawks	0.1058	0.0808	−0.0250	−23.6%
883	heineken	0.6273	0.6732	+0.0459	7.3%
884	qualcomm	0.4833	0.4959	+0.0126	2.6%
885	shimano	0.2291	0.4913	+0.2622	114.4%
886	west wing	0.3650	0.3782	+0.0132	3.6%
887	world trade organization	0.2459	0.2373	−0.0086	−3.5%
888	audi	0.5723	0.6080	+0.0357	6.2%
889	scientology	0.3115	0.3413	+0.0298	9.6%
890	olympics	0.2403	0.2547	+0.0144	6.0%
891	intel	0.0613	0.0662	+0.0049	8.0%
892	jim moran	0.5332	0.5316	−0.0016	−0.3%
893	zyrtec	0.0988	0.2045	+0.1057	107.0%
894	board chess	0.1696	0.2343	+0.0647	38.1%
895	oprah	0.1414	0.1732	+0.0318	22.5%
896	global warming	0.1418	0.1409	−0.0009	−0.6%
897	ariel sharon	0.0953	0.1121	+0.0168	17.6%
898	business intelligence resources	0.0074	0.0071	−0.0003	−4.1%
899	cholesterol	0.0385	0.0474	+0.0089	23.1%
900	mcdonalds	0.2289	0.2290	+0.0001	0.0%
all		0.3022	0.3259	+0.0237	7.8%

Products	Organizations	Politics
march of the penguins, macbook pro, blackberry, brokeback mountain, heineken, shimano, audi, intel, zyrtec	whole foods, sonic food industry, qualcomm, world trade organization, mcdonalds	larry summers, state of the union, ann coulter, abramoff bush, jon stewart, letting india into the club?, colbert report, basque, cheney hunting, joint strike fighter, muhammad cartoon, cindy sheehan, bruce bartlett, coretta scott king, jihad, west wing, jim moran, ariel sharon

Figure 7. Three categories of topics.

turned out to be effective for identifying opinions on them as well. This result suggests that the language model learned from these reviews are generalizable to products in general and even some types of non-products.

Next, regarding the topics for which retrieval performance decreased:

- **Ann Coulter** (#854) is an American conservative political commentator, syndicated columnist, and best-selling author.
- **Letting india into the club?** (#859).
- **Basque** (#865).
- **Jim Moran** (#892) has represented the 8th congressional district of Virginia since 1991. He is a member of the Democratic Party.
- **Fox news report** (#881) Fox News is a cable and satellite television news channel [...]. Some critics have asserted that both Fox's news reporting and its political commentary promote conservative political positions.
- **Sonic food industry** (#877) was not found in Wikipedia but refers to an American fast-food restaurant chain, Sonic Drive-In.
- **World Trade Organization** (#887) is an international organization designed to supervise and liberalize international trade.
- **Barry Bonds** (#870) Barry Bonds is a former Major League Baseball outfielder.
- **Brokeback Montain** (#872) Brokeback Mountain is a 2005 romantic drama film that depicts the complex romantic and sexual relationship between two men in the American West from 1963 to 1983.
- **Seahawks** (#882) Seahawks are a professional American football team based in Seattle, Washington.

Judging from their descriptions, the first five can be categorized as "politics" and the next two as "organizations" (and the rest as other miscellaneous categories). These categories of topics appear to be difficult to improve by using our language model, although the negative impact was relatively small.

Then, we grouped the 50 topics in Table 6 into the three general categories, i.e., products, politics, and organizations, to see if any difference with respect to the performance improvement/reduction exists across the different categories. Note that topics not under these categories were disregarded in this analysis. Figure 7 shows the categories of topics.

When examined per category, the "products", "politics", and "organizations" categories gained 20.1%, 0.1%, and 3.3% improvement, respectively, which emphasizes the difficulty of the latter two categories. Similar results (47.9%, 1.4%, and 7.6%, respectively) were observed when LM was used for initial retrieval. Taken altogether,

Table 7: Results of model adaptation using different triggering words

Configuration		MAP	Imprv. over Initial search
Initial search		0.1965	—
Reranking by trigger model		0.2398	22.0%
Adapted by	1) topic only	0.2430	23.6%
	2) pronouns only	0.2456*	25.0%
	3) topic + pronouns	0.2452*	24.8%

these results suggest that there are different vocabularies and/or trigger pairs used to express subjective opinions on politics or organizations from those found in product reviews. It should be stressed, however, that there are also many topics in the politics or organizations categories which were improved by applying our model, such as State of the union (+10.7%), Joint strike fighter (+30.7%), Bruce Bartlett (+10.6%), Jihad (+14.0%), and Qualcomm (+2.6%).

The next section applies the model adaptation technique to examine if further/any improvement is achieved.

Effectiveness of Model Adaptation

Based on the steps for language model adaptation described in the earlier section, we conducted additional experiments for opinion retrieval using model adaptation for the 50 topics from the 2006 Blog track. We experimentally used the top 50 blog posts from the initially retrieved set ($k = 50$). Note that only LM was used for this experiment because the indexed contents of the retrieved set were needed to apply model adaptation, which is highly dependent on the implementation of content extraction. Specifically, baseline4 is based on "cleaned" blog pages mostly containing only the blog posts themselves (Lee et al., 2008), whereas our baseline (and our language model) is built on noisier blog posts including all text in the pages.

Table 7 compares the results from previous experiments and those by the adapted trigger models, where the following three types of triggering words were tested: 1) only given topic titles, 2) only the predefined set of pronouns, and 3) both topic titles and pronouns. An asterisk indicates statistically significant improvement at the $p < 0.01$ level by sign test over the subjective trigger model without adaptation.

Overall, the performance in MAP more or less improved by adapting the trigger model to given topics irrespective of the types of triggering words considered. In particular, considering only pronouns lead to the highest improvement. Looking into individual topics (not shown), the most significant improvement was obtained for "Zyrtec" whose average precision jumped from 0.2187 by non-adaptation to 0.3230 (+47.7%), whereas the worst case was "Basque" whose average precision dropped from 0.2061 to 0.1673 (−18.8%). To highlight the difference between these two extremes, Table 8 presents some of the newly identified trigger pairs in descendant order of the number of times the particular trigger pairs were referenced in estimating adapted $P_E(\cdot)$. In this sense, they can be seen as most influential trigger pairs that contributed to the performance improvement/reduction.

In Table 8, we can observe that there are some trigger pairs deemed useful for Zyrtec, including "`i` → `sure`", "`it` → `perfect`", and "`i` → `bet`", whereas no such triggers can be recognized for Basque. New triggers identified

Table 8: Newly identified trigger pairs most often used for estimating $P_E(\cdot)$ for the topics "Zyrtec" and "Basque", where the numbers in parentheses indicate how many times the respective association $\alpha(b|a)$ was referenced to calculate $P_E(\cdot)$

Zyrtec			Basque		
you	→ year	(744)	this	→ spanish	(224)
i	→ case	(697)	you	→ come	(161)
it	→ case	(576)	i	→ spanish	(138)
you	→ d	(525)	i	→ told	(108)
i	→ sure	(516)	it	→ last	(97)
this	→ case	(495)	i	→ simply	(85)
it	→ perfect	(478)	this	→ city	(84)
i	→ bet	(456)	it	→ spanish	(78)
you	→ come	(418)	my	→ city	(70)
it	→ kind	(400)	this	→ road	(66)
my	→ year	(353)	i	→ city	(66)
i	→ extreme	(339)	i	→ south	(60)

for Basque are mainly general and geographical regarding the Spanish autonomous Basque region, including "this → spanish" and "i → south". Different from these triggers, the user intention for this query presupposed by the Blog track was about political movement advocating for independence by Basque nationalism according to its topic definition.[7] This intention is difficult to perceive only from the one-word query, "basque". It suggests that the adaptation of a subjective trigger model may not be effective or even hazardous when user information needs are not evident from a user query, such as this topic.

Then, we looked at difficult individual topics, Ann Coulter, Cindy Sheehan, Sonic food industry, West Wing, World Trade Organization, and Jim Moran, for which reranking by our (non-adapted) subjective trigger model decreased retrieval performance. This experiment will reveal if there is any positive effect of newly identified trigger pairs on these difficult topics. Table 9 shows their average precision scores and percent improvement as compared with those by the original non-adapted subjective trigger model (denoted as "Trigger" in the table). As can be seen, most topics showed more or less positive results through the model adaptation, even though the effect is limited and only two of them ("Ann Coulter" and "sonic food industry") exceeded the *initial* retrieval performance.

In addition, we conducted the model adaptation experiment on the topic sets from the 2007 and 2008 Blog tracks using the predefined set of pronouns as triggers. Similar to the above results, we observed marginal increase in MAP from 0.3280 to 0.3363 for the 2007 topic set (+2.5%), and 0.3057 to 0.3075 for the 2008 topic set (+0.6%). The reason why the model adaptation worked only marginally may be due to the use of the top k blog posts initially retrieved. For the adaptation to work better, the k posts should ideally be all subjective and contain opinions about only the target in question, which is unlikely. Focusing only on subjective sentences about only the target may make the model adaptation more effective. Also, collecting more opinionated sentences, for example, from web search

[7] When the topic title "basque" was used as the triggering word, more relevant words, such as "right" and "liberal", were discovered as triggered words. Still, these trigger pairs did not improve the non-adapted result.

Table 9: Performance (average precision) change for difficult topics before/after model adaptation

Topic #	Topic	Trigger	Adapted	% imprv.
854	ann coulter	0.4591	0.4838	+2.5%
871	cindy sheehan	0.4576	0.4640	+0.6%
877	sonic food industry	0.0380	0.0453	+0.7%
886	west wing	0.2407	0.2410	+0.0%
887	world trade organization	0.0658	0.0653	−0.1%
892	jim moran	0.4728	0.4891	+1.6%

results may be beneficial, too. We will continue to study better use of trigger pairs for opinion retrieval.

Polarized opinion retrieval

Overview

We have so far focused on opinion retrieval disregarding the polarity of opinions (i.e., positive and negative). However, it may be beneficial to distinguish positive and negative opinions if, for example, we are specifically looking for pros and/or cons of the target of the interest. Such scenarios occur frequently when we are to make decisions, such as whether to buy a particular product, whether to stay at a particular hotel, and so on. This problem, called the "polarity task", was also tackled at the Blog track from 2007 to 2008. Note that the relevance judgment for the year 2006 also includes polarity information, which allowed us to use the topics for 2006 as well for the following evaluation. This section conducts additional experiments to see whether or not our proposed subjective trigger model is also effective for polarized opinion retrieval.

Experimental settings

As we have described, we initially developed the subjective trigger model for opinion retrieval. However, it is straightforward to apply exactly the same model to the polarity task. What is needed are a corpus of positive opinions and another corpus of negative opinions, which are readily available at `Amazon.com`. Instead of feeding Amazon customer reviews as a whole to build a single subjective trigger model as it has been done, we can separately build a model using only positive reviews, say with five stars, and another model using only negative reviews, say with one star. Then, with the same framework as the opinion retrieval, these models could be used to identify blogs containing positive and negative opinions, respectively.

Along this line, we carried out experiments for the polarity task as follows. We acquired 5,000 customer reviews from Amazon for each of five stars and one star and then created a subjective trigger model for each as described in the *Building a Subjective Trigger Model* section. For initial search, we adopted the same LM as has been used and additionally baseline4 for comparison. The performance of polarized opinion retrieval by the initial search models is shown in Table 10 (before reranking by the subjective trigger models), where we looked at the topic sets for the 2006 to 2008 Blog tracks and used MAP as the evaluation metric, which is the same metric as used at the 2008 Blog track. For positive opinion retrieval, the blog posts with the label "4" (i.e., topically relevant and only positively

Table 10: Performance of polarized opinion retrieval in MAP by initial search models (before reranking)

Model	Polarity	MAP			
		2006	2007	2008	All
baseline4	Positive	0.1080	0.2176	0.1563	0.1607
	Negative	0.0922	0.0925	0.1284	0.1058
	Average	0.1001	0.1551	0.1424	0.1333
LM	Positive	0.0672	0.1392	0.0947	0.1004
	Negative	0.0673	0.0511	0.0910	0.0695
	Average	0.0673	0.0952	0.0929	0.0850

opinionated) were treated as relevant, and for negative opinion retrieval, those with "2" (i.e., topically relevant and only negatively opinionated) were treated as relevant. Blog posts with labels "3" (i.e., topically relevant and both positively and negatively opinionated) were not regarded as relevant for either case, following the evaluation methodology for the 2008 Blog track. As with the opinion retrieval task, baseline4 is again the strongest baseline among the results submitted by the 2008 Blog track participants (Ounis et al., 2008).

Results

First, let us look at the two sets of trigger pairs independently identified for positive and negative opinions. Table 11 contrasts positively and negatively subjective trigger pairs identified by the modified low-level trigger criterion by using the one-star and five-stars Amazon customer reviews, respectively.

Many of the prominent trigger pairs overlap with those in Table 2 identified on Amazon reviews (which is not surprising). We can see, however, that there is a mild separation between the positive and negative sets and some triggers found as positively subjective triggers are quite explicit, such as "`i` → `loved`" and "`I` → `liked`", although there are some trigger pairs equally used in both directions of opinions (e.g., "`I` → `couldn`"). Overall, it seems that negative opinions tend to be expressed more indirectly and politely. Some representative examples found in the reviews include "**I guess** it might be worth your reading if you need ..." and "**I wish** I had gotten [it] at the library" (implying she should not have bought it by herself), and our approach successfully captures these subtle triggers shown in bold in the examples. Note that, more direct expressions (e.g., "`I` → `hate`") were also identified as negatively subjective triggers (not shown in the short list).

Using these subjective trigger models, we reranked the initial search results based on Equation (11) for finding positive and negative opinions, respectively. The results are summarized in Table 12.

As can be seen, the retrieval performance improved in all the cases, irrespective of the polarities of the opinions, topic sets (i.e., 2006, 2007, or 2008), or the initial retrieval models, although the degree of the improvement differs depending on different settings. When LM was used for initial retrieval, the overall improvement was 18.6%, whereas the stronger baseline, baseline4, is much harder to improve; achieving only 4.17% increase in MAP on average. It should be mentioned that, however, there were only three groups (excluding ours) out of 11 who used baseline4 managed to improve over baseline4 at the 2008 Blog track. These groups utilized a variety of evidence to

Table 11: Most prominent subjective triggers identified by the modified criterion using 5000 five-stars or one-star Amazon customer reviews

Positively subjective triggers				Negatively subjective triggers			
Triggering (a)		Triggered (b)	$\Delta_{a\to b}$	Triggering (a)		Triggered (b)	$\Delta_{a\to b}$
this	→	be	5.133	i	→	guess	5.156
i	→	loved	4.958	i	→	wish	5.150
this	→	all	4.889	this	→	read	4.985
this	→	your	4.799	i	→	felt	4.822
you	→	your	4.795	i	→	agree	4.770
i	→	couldn	4.766	i	→	suppose	4.753
i	→	cannot	4.638	i	→	purchased	4.687
this	→	she	4.568	i	→	kept	4.664
i	→	hope	4.495	this	→	more	4.663
this	→	they	4.492	i	→	more	4.655
this	→	gives	4.477	i	→	no	4.622
i	→	learned	4.477	i	→	couldn	4.577
i	→	they	4.458	this	→	no	4.564
this	→	helped	4.417	i	→	personally	4.561
i	→	liked	4.413	i	→	mean	4.512
i	→	d	4.413	this	→	wish	4.495
...				...			

Table 12: Performance of polarized opinion retrieval in MAP after reranking by positively/negatively subjective trigger models. Percentages in parentheses indicate performance increase with respect to the initial results before reranking (upper rows)

Model	Polarity		MAP			
			2006	2007	2008	All
baseline4	Positive	Before	0.1080	0.2176	0.1563	0.1607
		After	0.1115 (+3.24%)	0.2234 (+2.67%)	0.1602 (+2.50%)	0.1651 (+2.74%)
	Negative	Before	0.0922	0.0925	0.1284	0.1058
		After	0.0970 (+5.21%)	0.1019 (+10.16%)	0.1350 (+5.14%)	0.1125 (+6.33%)
	Average	Before	0.1001	0.1551	0.1424	0.1333
		After	0.1043 (+4.15%)	0.1627 (+4.90%)	0.1476 (+3.69%)	0.1388 (+4.17%)
LM	Positive	Before	0.0672	0.1392	0.0947	0.1004
		After	0.0767 (+14.14%)	0.1662 (+19.40%)	0.1112 (+17.42%)	0.1181 (+17.63%)
	Negative	Before	0.0673	0.0511	0.0910	0.0695
		After	0.0756 (+12.33%)	0.0721 (+41.10%)	0.1008 (+10.77%)	0.0834 (+20.00%)
	Average	Before	0.0673	0.0952	0.0929	0.0850
		After	0.0762 (+13.23%)	0.1192 (+25.22%)	0.1060 (+14.16%)	0.1008 (+18.60%)

identify the polarity of opinionated blog pages. For example, Bermingham et al. (2008) looked at document length, unusual word/punctuation patterns (e.g., "arrrrgh") expressed as regular expressions, part-of-speech *n*-grams, Penn Treebank phrasal types (e.g., NP and AJDP), and vocabularies from SentiWordNet (Esuli & Sebastiani, 2007). Yang (2008) used Wilson's subjective terms (Wiebe et al., 2004), a subjective lexicon constructed from the Blog 06 corpus and the Internet Movie Database (IMDb) reviews, and term collocations involving first and second person pronouns (e.g., I and you). Unusual word/punctuation patterns mined from infrequent terms were used as well. Moreover, he took into account valence shifters and reversed the default polarity associated with a term if it appears near a negative term, such as "not", "never", "no", and "without".

Related Work

Reflecting the intense interest in blogs or UGC in general, there is a large body of research conducted for opinion mining and sentiment analysis. Among them, opinionated document retrieval or opinion retrieval is specifically focusing on retrieving documents containing subjective opinions on the the target described as a user query. Opinion retrieval has been widely studied by many research groups in the last four years, partly motivated by the Blog track (Ounis et al., 2006) introduced to TREC in 2006. In the track, opinion retrieval was tackled as one of the shared task challenges from 2006 and 2008. Other tasks include, in addition to polarized opinion retrieval discussed above, blog finding distillation (Elsas et al., 2008), faceted blog distillation, and top stories identification tasks.[8]

For opinion retrieval, whether polarized or not, most participants adopted a two-tier framework as with this study; they first conducted an initial search for locating topically relevant blog posts and then applied a variety of techniques to identify opinionated posts within the initial retrieval set. The latter, opinion-specific features can be roughly divided into two approaches, namely, supervised classification-based and lexicon-based.

The first type of approaches used supervised classifiers to judge if a retrieved blog post is a subjective opinion (Gerani et al., 2009; Seki et al., 2007; Zhang & Yu, 2006; Zhang et al., 2007). For example, Zhang et al. (2007) collected a large number of opinionated and non-opinionated documents from the web, specifically, RateItAll[9] for opinionated documents and Wikipedia for non-opinionated, to train a per-topic classifier with word unigrams and bigrams as features. The classifier was applied to each sentence ($\in$ blog post d) containing query terms and the classification results were aggregated to measure the overall opinionatedness of d. Their reported best MAP for opinion retrieval on the 2006 topic set is 0.2726. Note that the result cannot be directly compared with those reported in this paper (0.2398 with LM and 0.3259 with baseline4) as their initial search model for retrieving topically relevant blog posts was different from our study.

The second, more popular type of approaches are lexicon-based, which automatically or manually construct a subjective word/phrase list and use it for estimating the opinionatedness of a given blog post (Hannah et al., 2007; He et al., 2008; Lee et al., 2008; Mishne, 2006; Oard et al., 2006; Vechtomova, 2010; Yang et al., 2006; Zhang & Ye, 2008). For example, Hannah et al. (2007) compiled an English word list from various linguistic resources, such as OpinionFinder (Wilson et al., 2005), and computed for each word the opinionated discriminability based on the

[8] http://ir.dcs.gla.ac.uk/wiki/TREC-BLOG

[9] http://www.rateitall.com/

relevance judgment from the 2006 Blog track opinion retrieval task. The weighted word list was then used as a query submitted to an IR system in order to measure the opinionated nature of each document; the relevance scores returned by the IR system were considered as opinion scores for the documents and were used for reranking initial retrieval results as with our study. Their approach was reported effective, having yielded the highest improvement of 15.87% in MAP over their initial search at the TREC 2007 Blog track. He et al. (2008) extended Hannah et al. (2007)'s work and used, instead of external linguistic resources, the Blog06 corpus to compile an English word dictionary based on the skewed query model (Cacheda et al., 2005), although using external resources was found more effective. Another approach in the lexicon-based category was proposed by Lee et al. (2008). For initial retrieval of topically relevant blog posts, they looked at both document-level and passage-level relevance and combined them by way of linear interpolation, having produced the strongest baseline, baseline4. For estimating opinionatedness of a blog post, they first estimated the subjectivity of word w, represented as a probability $P(Sub|w)$, and added up the probabilities of words composing a blog post. The probability $P(Sub|w)$ was estimated using SentiWordNet (Esuli & Sebastiani, 2007) and Amazon customer reviews. Finally, the topically relevance scores and opinionatedness scores were combined by linear interpolation. They have shown the strongest performance in the literature; a MAP of 0.4061 over all the topic sets as compared with 0.3791 produced by our approach using the same baseline4.[10]

As compared with these lexicon-based approaches, an obvious advantage of our proposed approach is that our approach does not require any precompiled lexicons or thesauri. Also, our approach does not require any relevance judgment data as used by He et al. (2008) but only a corpus of opinions. While relevance judgment data are usually not available, such a corpus of opinion is abundant at customer review sites like `Amazon.com`. Even though our approach does not yield the best performance, these properties, together with the simple framework, make our approach particularly attractive for porting to other languages as long as the same assumptions apply (i.e., an opinion is composed of two constituents, one is the subject or object and the other is a subjective expression).

Overall, our proposed approach is novel in a sense that it does not belong to either category summarized above. To the best of our knowledge, none has attempted to capture long-distance dependencies focused on subjective opinions by way of language modeling and successfully applied it to opinion retrieval.

Conclusions and Future Work

This paper discussed an application of a focused trigger model to opinion retrieval by way of reranking initial search results. Evaluative experiments on the TREC Blog track test collections showed that by incorporating subjective trigger pairs, retrieval performance increased by 5.6% to 33.4% in MAP depending on the initial retrieval models and data sets. Also, a closer analysis indicated that the identified triggers did capture the characteristics of opinions as compared with a baseline trigram model, contributing to discriminating opinionated posts from the non-opinionated. When looking at individual topics, it was found that there were some types of topics, specifically, politics and organizations, that are more difficult to improve by the proposed trigger model. To deal with it, we proposed a

[10] Lee et al. (2008) do not explicitly mention that they used baseline4 but we assume so as it was the best result they obtained.

mechanism to dynamically update the trigger model to a given topic, which overall had marginal but positive effects for most topic types. Furthermore, we demonstrated that the subjective trigger model was effective for polarized opinion retrieval as well by creating separate language models for positive and negative opinions, respectively.

Despite the effectiveness demonstrated in the experiments, our proposed model still has several limitations. One of them stems from our primary assumption that a subjective opinion is composed of two constituents, i.e., subject (or object) and a subjective expression, disregarding a target. This model may be suitable for general opinion retrieval but may not be so for targeted opinion retrieval for a given query or target, which was the main theme of the present work. Since trigger pairs do not explicitly consider the target, our model by design ignores the association between the target and a subjective expression. Our model adaptation using a given query as triggering words intended to partly address this issue but was found less effective than using pronouns alone. A better subjective trigger model may be created by focusing *only* on subjective sentences or clauses containing *only* the target in question. Also, our second assumption considered only pronouns as triggering words and consequently our subjective trigger model ignores many sentences not containing pronouns, such as "John likes . . .". In order for our model to learn from such examples, proper/common nouns need to be converted to appropriate pronouns (e.g., "he") in advance.

Another limitation is caused by the fact that our approach is word-based, which cannot properly handle phrases. For example, it is impossible to identify a trigger pair "it → piece of junk" in a sentence "it is a piece of junk" unless such phrases are recognized as one-words by, for example, noun chunking. Note that, however, this problem is partly dealt with by our n-gram model combined with a trigger model (our final model combines an n-gram and a trigger models; see Equation (1)). An n-gram model would learn word patters like "is a piece", "a piece of", and "piece of junk" if they appear in a corpus of opinions and thus give higher probabilities (opinionatedness) to text containing such patterns. Another problem for a word-based approach is that it cannot understand syntactic structure, such as the subjunctive mode. For instance, our model would simply take "it → great" in "it would've been a great purchase if it wouldn't have broken" as an opinion cue without considering the condition "if it wouldn't [. . .]". Furthermore, when applied to polarized opinion retrieval, our word-based approach cannot deal with valence shifters, which are crucial to determine opinion's polarity. Within the current framework, however, such information could be encoded by adding a special prefix (e.g., "UN-") to subjective expressions. For example, a positive opinion "this unit is feature rich without being too complicated to use" may produce trigger pairs, such as "this → rich" and "this → complicated". The latter implies negativity but can be modified to "this → UN-complicated" if the valence shifter "without" is considered. These valence shifters can be detected by syntactic analysis or heuristic rules similar to those used by Yang (2008).

For future work, we would like to explore alternative textual resources, such as the Blog06 test collection (He et al., 2008) and the Web, for language modeling. Also, we are interested in investigating better representation of blog posts; our current framework treats each blog post as a long sequence of words, which would contain many words irrelevant to a given topic. A common window- or passage-based approach (Santos et al., 2009; Lee et al., 2008) taking words around the topic would be beneficial. Lastly, since our approach is different from other classification- and lexicon-based approaches, it could be possible to combine it with these types of approaches for further improvement.

As a final remark, we hope that our work would encourage novel applications of statistical NLP techniques to information retrieval. Language model-based approaches (Lafferty & Zhai, 2001; Ponte & Croft, 1998) have been around in IR but they are typically based on uni-grams (bag-of-words), although much work has been done to incorporate term dependencies (e.g., Metzler & Croft, 2005). There are many statistical NLP models that have not yet been explored for IR, including trigger models, to look at wider and richer context than single words. They could be useful for practical IR problems, such as opinion retrieval, which are better addressed by considering word associations, sentence structures, and meanings. We will continue to work on the integration of NLP and IR for better information access.

References

Adar, E., & Adamic, L. (2005). Tracking information epidemics in blogspace. In *Proceedings of the 2005 IEEE/WIC/ACM international conference on web intelligence* (pp. 207–214).

Agarwal, N., Liu, H., Tang, L., & Yu, P. S. (2008). Identifying the influential bloggers in a community. In *Proceedings of the international conference on web search and web data mining* (pp. 207–218).

Bermingham, A., Smeaton, A., Foster, J., & Hogan, D. (2008). DCU at the TREC 2008 blog track. In *Proceedings of the 17th text retrieval conference (TREC 2008).*

Cacheda, F., Plachouras, V., & Ounis, I. (2005). A case study of distributed information retrieval architectures to index one terabyte of text. *Information Processing & Management*, *41*(5), 1141–1161.

Ding, X., Liu, B., & Yu, P. S. (2008). A holistic lexicon-based approach to opinion mining. In *Proceedings of the international conference on web search and web data mining* (pp. 231–240).

Elsas, J. L., Arguello, J., Callan, J., & Carbonell, J. G. (2008). Retrieval and feedback models for blog feed search. In *Proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval* (pp. 347–354).

Esuli, A., & Sebastiani, F. (2007). PageRanking WordNet synsets: An application to opinion mining. In *Proceedings of the 45th annual meeting of the association for computational linguistics.*

Gerani, S., Carman, M., & Crestani, F. (2009). Investigating learning approaches for blog post opinion retrieval. In *Proceedings of the 31st european conference on information retrieval (ecir09)* (pp. 313–324).

Hannah, D., Macdonald, C., Peng, J., He, B., & Ounis, I. (2007). University of Glasgow at TREC 2007: Experiments in blog and enterprise tracks with Terrier. In *Proceedings of the 16th text retrieval conference.*

He, B., Macdonald, C., He, J., & Ounis, I. (2008). An effective statistical approach to blog post opinion retrieval. In *Proceeding of the 17th acm conference on information and knowledge management* (pp. 1063–1072).

Jindal, N., & Liu, B. (2008). Opinion spam and analysis. In *Proceedings of the international conference on web search and web data mining* (pp. 219–230).

Jurafsky, D., & Martin, J. H. (2000). *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition* (S. Russell & P. Norvig, Eds.). Prentice Hall.

Lafferty, J., & Zhai, C. (2001). Document language models, query models, and risk minimization for information retrieval. In *Proceedings of the 24th annual international acm sigir conference on research and development in information retrieval* (pp. 111–119).

Lau, R., Rosenfeld, R., & Roukos, S. (1993). Trigger-based language models: a maximum entropy approach. In *Proceedings of ieee international conference on acoustics, speech, and signal processing* (Vol. 2, pp. 45–48).

Lavrenko, V., & Croft, B. (2001). Relevance based language models. In *Proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 120–127).

Lee, Y., Na, S.-H., Kim, J., Nam, S.-H., Jung, H. young, & Lee, J.-H. (2008). KLE at TREC 2008 blog track: Blog post and feed retrieval. In *Proceedings of the 17th text retrieval conference (TREC 2008).*

Macdonald, C., Ounis, I., & Soboroff, I. (2007). Overview of the TREC-2007 blog track. In *Proceedings of the 16th text retrieval conference.*

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT Press.

Mei, Q., Ling, X., Wondra, M., Su, H., & Zhai, C. (2007). Topic sentiment mixture: Modeling facets and opinions in weblogs. In *Proceedings of the 16th international world wide web conference.*

Metzler, D., & Croft, W. (2004). Combining the language model and inference network approaches to retrieval. *Information Processing and Management. Special Issue on Bayesian Networks and Information Retrieval*, *40*(5), 735–750.

Metzler, D., & Croft, W. B. (2005). A markov random field model for term dependencies. In *Proceedings of the 28th annual international acm sigir conference on research and development in information retrieval* (pp. 472–479).

Mishne, G. (2006). Multiple ranking strategies for opinion retrieval in blogs. In *Proceedings of the 15th text retrieval conference.*

Oard, D., Elsayed, T., Wang, J., Wu, Y., Zhang, P., Abels, E., Lin, J., & Soergel, D. (2006). TREC-2006 at Maryland: Blog, enterprise, legal and QA tracks. In *Proceedings of the 15th text retrieval conference.*

Ounis, I., Macdonald, C., & Soboroff, I. (2008). Overview of the TREC 2008 blog track. In *Proceedings of the 17th text retrieval conference (TREC 2008).*

Ounis, I., Rijke, M. de, Macdonald, C., Mishne, G., & Soboroff, I. (2006). Overview of the TREC-2006 blog track. In *Proceedings of the 15th text retrieval conference.*

Ponte, J. M., & Croft, W. B. (1998). A language modeling approach to information retrieval. In *Proceedings of the 21st annual international acm sigir conference on research and development in information retrieval* (pp. 275–281).

Sakai, T., Manabe, T., & Koyama, M. (2005). Flexible pseudo-relevance feedback via selective sampling. *ACM Transactions on Asian Language Information Processing*, *4*(2), 111–135.

Salton, G., & McGill, M. J. (1983). *Introduction to modern information retrieval*. McGraw-Hill, Inc.

Santos, R. L., He, B., Macdonald, C., & Ounis, I. (2009). Integrating proximity to subjective sentences for blog opinion retrieval. In *Proceedings of the 31st european conference on information retrieval (ecir09)* (pp. 325–336).

Seki, K., Kino, Y., Sato, S., & Uehara, K. (2007). TREC 2007 blog track experiments at Kobe University. In *Proceedings of the 16th text retrieval conference.*

Sparck Jones, K. (1972). Statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, *28*(1), 11–20.

Tao, T., & Zhai, C. (2006). Regularized estimation of mixture models for robust pseudo-relevance feedback. In *Proceedings of the 29th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 162–169).

Tillmann, C., & Ney, H. (1996). Selection criteria for word trigger pairs in language modeling. In *Proceedings of the 3rd international colloquium on grammatical inference* (pp. 95–106).

Vechtomova, O. (2010). Facet-based opinion retrieval from blogs. *Information Processing & Management*, *46*(1), 71–88.

Voorhees, E. M., & Harman, D. K. (Eds.). (2005). *Trec: Experiment and evaluation in information retrieval*. The MIT Press.

Wiebe, J., Wilson, T., Bruce, R., Bell, M., & Martin, M. (2004, September). Learning subjective language. *Computational Linguistics*, *30*, 277–308.

Wilson, T., Hoffmann, P., Somasundaran, S., Kessler, J., Wiebe, J., Choi, Y., Cardie, C., Riloff, E., & Patwardhan, S. (2005). OpinionFinder: a system for subjectivity analysis. In *Proceedings of hlt/emnlp on interactive demonstrations* (pp. 34–35).

Yang, K. (2008). WIDIT in TREC-2008 blog track: Leveraging multiple sources of opinion evidence. In *Proceedings of the 17th text retrieval conference.*

Yang, K., Yu, N., Valerio, A., & Zhang, H. (2006). WIDIT in trec-2006 blog track. In *Proceedings of the 15th text retrieval conference.*

Yang, K., Yu, N., & Zhang, H. (2007). WIDIT in TREC 2007 blog track: Combining lexicon-based methods to detect opinionated blogs. In *Proceedings of the 16th text retrieval conference.*

Zhang, M., & Ye, X. (2008). A generation model to unify topic relevance and lexicon-based sentiment for opinion retrieval. In *Proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval* (pp. 411–418).

Zhang, W., & Yu, C. (2006). UIC at TREC 2006 blog track. In *Proceedings of the 15th text retrieval conference.*

Zhang, W., Yu, C., & Meng, W. (2007). Opinion retrieval from blogs. In *Proceedings of the sixteenth acm conference on conference on information and knowledge management* (pp. 831–840).

Zhou, G., Joshi, H., & Bayrak, C. (2007). Topic categorization for relevancy and opinion detection. In *Proceedings of the 16th text retrieval conference.*