Adaptive Subjective Triggers for Opinionated Document Retrieval

Kazuhiro Seki
Organization of Advanced Science & Technology
Kobe University
1-1 Rokkodai, Nada, Kobe 657-8501, Japan
seki@cs.kobe-u.ac.jp

Kuniaki Uehara
Graduate School of Engineering
Kobe University
1-1 Rokkodai, Nada, Kobe 657-8501, Japan
uehara@kobe-u.ac.jp

ABSTRACT

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 by over 20% in MAP. 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.

Categories and Subject Descriptors

H.3.3 [**Information storage and retrieval**]: Information Search and Retrieval—*selection process*; I.2.7 [**Artificial intelligence**]: Natural Language Processing—*language models, text analysis*

General Terms

Algorithm, Experimentation, Languages

Keywords

Weblog, opinion retrieval, trigger language models

1. 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 recently 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 NLP, information retrieval (IR), and other research communities as an attractive domain for exploration [1, 2, 3, 5, 14].

Among a variety of research opportunities targeting blogs, this paper focuses on IR aspects, specifically, opinionated document (blog post) retrieval, which has been challenged at the Text Retrieval Conference (TREC) Blog Track since 2006 [12, 18]. The approaches explored by the track participants and others can be roughly categorized into lexicon-based [16, 17, 29] and classification-based [21, 30, 31]. 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 opinionatedness. The latter, classification-based, also (typically) relies on word occurrences but automatically create a classifier based on positive (opinionated) and negative (non-opinionated) examples using machine learning techniques.

In this paper, we propose a novel and effective approach to opinionated document retrieval (or opinion retrieval for short) which does not belong to either category. Our approach was partly inspired by the empirical finding that considering the proximity of pronouns and subjective expressions to objects improves opinion retrieval [32]. 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 trigger model [9, 25], 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 or the object that the opinion is about, and the other is a subjective expression. 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.

The rest of this paper is structured as follows: Section 2 details our approach to building a trigger model for subjective opinions. Section 3 evaluates the validity of our proposed model and its effectiveness in retrieving opinionated blog posts. Section 4 summarizes the related work. Lastly, Section 5 concludes with a brief summary of the findings and possible future directions.

WSDM'09 February 9–12, 2009, Barcelona, Spain

2. OPINION RETRIEVAL BY A TRIGGER MODEL

2.1 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 [16, 17, 27]. 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." might possibly be an opinion but is rather objective. To distinguish such difference, one may 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 [13], 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 word. 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 practical reasons, n is often set to 2 or 3 depending on the intended application and 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.

2.2 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. [9] proposed the trigger-based language model (or trigger model for short). A trigger refers to a pair of words, one tends to bring about the occurrence of the other. A trigger model $P_T(w|h)$ incorporating such trigger pairs is used to enhance a baseline n-gram language model $P_B(w|h)$ by linearly interpolating the two:

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

where w and h denote a word and a history (a set of words preceding w), respectively, and λ is the interpolation parameter. (We briefly describe the definition of $P_T(w|h)$ in the end of Section 2.3.)

To build a trigger model, we first need to identify significant triggering and triggered word pairs. Given a corpus of documents, any word pair in the vocabularies can potentially be a trigger pair. Tillmann and Ney [25] 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,

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

Each word b satisfying Equation (2) is evaluated in combination with every word a in the vocabulary 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 text represented as a word sequence $w_1, w_2, \ldots, w_m$, the difference $\Delta_{a\rightarrow b}$ is computed as follows.

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

The better the extended model predicts the input text, the greater the difference becomes. After evaluating each word pair, 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*.

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 typically 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 [28, 32].

As compared to the ad hoc heuristics used in the previous work (see Section 4), 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*.

2.3 Building a Subjective Trigger Model

Based on the procedure and assumptions described in the previous section, we 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`. These reviews are written for various kinds of products sold at Amazon, including books, DVDs, electrical appliances, toys, etc. Their customer ratings (ranging from 1 to 5) were not distinguished in this study 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 up to 10,000 trigger pairs using the low level triggers criterion. In building the model, we limited the history h to the prior context (preceding words) in the same sentence. Table 1 shows the identified trigger pairs with highest log-likelihood improvement.

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 lower $P(w|h)$ than threshold t (see Equation (2)). 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,

Table 1: Most prominent low level triggers

Triggering (a)		Triggered (b)	$\Delta_{a \to b}$
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
		...	

we modified the criterion as follows:

$$\tau \cdot P(w_i|h_i) < t \tag{4}$$

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 → an" as in "This work does an outstanding job [...]".

Table 2 shows some trigger pairs with highest log-likelihood improvement using the modified criterion. 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, such as "this → an" as in the above example. 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". Although a similar set of word pairs could be identified by employing a syntactic parser, our proposed framework have two advantages over such approaches. 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 similar assumptions apply. Second, since it 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".

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
		...	

For each identified trigger pair ($a \to b$), their association, $\alpha(b|a)$, is calculated as normalized maximum likelihood estimate based on their collocations, which is in turn used to build the trigger model as follows.

$$P_T(w|h) = \frac{1}{|h|} \sum_{w_j \in h} \alpha(w|w_j) \tag{5}$$

Equation (5) 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 and Ney [25].

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 higher weight to the trigger model.

2.4 Model Adaptation

Since the subjective trigger model was built on Amazon customer reviews, which essentially deal with only 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) [10, 19, 24] which expands an 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 method.

1. Carry out initial search by a choice of an IR model for a given topic I.
2. Among the top k blog posts retrieved, identify trigger pairs, $a \to b$, and compute their associations, $\alpha'(b|a)$ in the same way described in Section 2.2. In this step, one could use the given topic as potential triggering words in addition to the predefined set of 14 pronouns (see Section 2.3).
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 (5), we use Equation (6).

$$P_T(w|h) = \frac{1}{|h|} \sum_{w_j \in h} \max\left(\alpha(w|w_j), \alpha'(w|w_j)\right) \tag{6}$$

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 reported in the literature [12].

3. EMPIRICAL EVALUATION

3.1 Data Set

To evaluate the validity of the proposed model, we conducted evaluative experiments on the Blog06 test collection provided for the TREC Blog Track 2006 [18]. It is a collection of over 3.2 million blog posts crawled over an 11 week period from December

2005 to February 2006. The collection also contains 50 topics, developed from commercial blog search engine query logs, describing user information needs. Figure 1 gives an example topic, where the "title" field indicates the actual query used by search engine users.

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 1: Example topic from the TREC 2006 Blog Track.

For each topic, relevant/irrelevant blog posts are marked in the collection 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 opinionated (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 Blog06 collection, we evaluated our proposed model in two ways. First, Section 3.2 assesses the validity of the language model itself out of the context of IR. Second, Section 3.3 examines the effectiveness of the model for opinion retrieval in an IR setting, followed by an evaluation of the model adaptation.

3.2 Evaluation of the Language Model

The subjective trigger model in Section 2.3 was created not on opinionated blogs but on Amazon customer reviews, which are reported to contain many "spam" reviews [7]. Therefore, we first examined the subjective trigger model whether it 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 [8]. Intuitively, perplexity 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 $2^{H(L,d)}$, where $H(L,d)$ denotes cross entropy of language model L on document d. Cross entropy is an information theoretic measure of distance between an estimated and true probability distributions and defined as in Equation (7) for large m.

$$\begin{aligned} H(L,d) &\approx -\frac{1}{m}\log P(w_1 \ldots w_m) \\ &\approx -\frac{1}{m}\sum_{i=1}^{m}\log P(w_i|h_i) \end{aligned} \tag{7}$$

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). Table 3 presents perplexity results for baseline language model P_B and 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

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%)

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. This observation is particularly important because it indicates that the subjective trigger pairs brought additional clues of opinionatedness that could not be captured by standard n-gram models.

The above experiment verifies 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 document level by comparing the distributions of cross entropy of P_E on opinionated and non-opinionated blog post, where n was set to 3. The result shown in Figure 2 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.

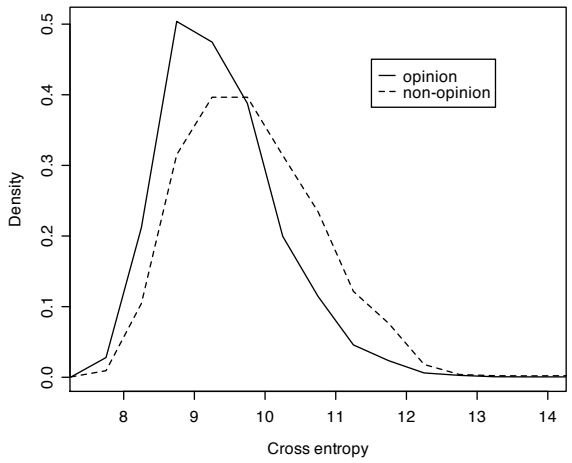

Figure 2: Distributions of cross entropy for opinionated and non-opinionated blog posts.

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.

3.3 Opinion Retrieval with the Subjective Trigger Model

3.3.1 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 [20] with the TFIDF term weighting [23], referred to as VSM, and b) the inference network model combined with a language modeling approach [15], referred to as INM. 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 1). Remember that we consider the labels 2–4 as relevant disregarding the polarity of opinions, i.e., positive, negative, and mixed (see Section 3.1). Table 4 shows the initial retrieval results in mean average precision (MAP) and also shows the TREC 2006 Blog Track official results (only using topic titles) and other results reported at post-TREC conferences for reference. Note that the TREC and post-TREC results were obtained by using a variety of opinion finding features, whereas VSM and INM are simply initial search results without activating any such functionalities.

Table 4: Initial search results using alternative IR models, where TREC 2006 official results and other post-TREC results are shown for reference

		MAP
Initial	VSM	0.1126
	INM	0.1965
TREC	Best	0.1885
	Median	0.1156
	Worst	0.0000
post-TREC	W. Zhang et al. [31]	0.2726
	M. Zhang and Ye [29]	0.2257

There is a large difference in performance between the two alternative IR models, VSM and INM, and INM even outperforms the best reported official MAP. Due to the observed advantage of INM,[1] the following experiments use only the INM result and attempt to improve the performance in terms of opinion retrieval.

3.3.2 Integration of a Subjective Trigger Model

In the initial retrieval by INM, 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:

$$Scr(d, I) = (1 - \beta) \cdot \log P(I|d) + \frac{\beta}{m} \sum_i \log P_E(w_i|h_i) \qquad (8)$$

where β is a interpolation parameter controlling the effect of the language model enhanced by subjective triggers. Notice that the second term corresponds to cross entropy in Equation (7). 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 [29].

We gradually increased the parameter β from 0 to 1 in Equation (8) and reranked the initially retrieved documents to see if any improvement in MAP is observed. Figure 3 shows the transition of the MAP score for different values of β. The leftmost circle, where $\beta = 0$, corresponds to the initial result. By varying β, MAP significantly 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 (8) is effective for spotting opinionated blog posts without degrading the initial topic-based ranking if β is properly chosen. Although not presented here, similar improvement from 0.2508 by initial retrieval by INM to 0.3072 (+22.5%) was observed for another set of 50 topics from the following TREC 2007 Blog Track, where β was again found optimal at around 0.35. This result implies the stability of optimum β across different topics.

It is worth mentioning that the result cannot be directly compared to those in Table 4 since each research group used different initial search module, which is reported to have a strong influence on the performance of opinion retrieval [11, 12]. To make the point, we applied the subjective trigger model to a stronger baseline[2] with a MAP of 0.3022 for initial search. Even with the quite strong baseline, our proposed approach managed to improve the performance by 0.3221 (+6.6%), which is the best MAP reported in the literature.

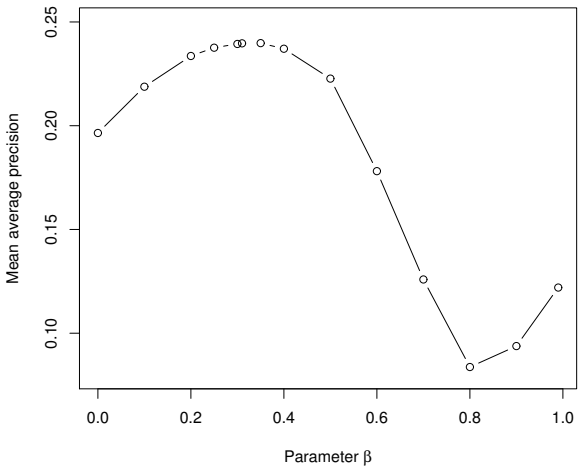

Figure 3: Relation between parameter β and MAP.

3.3.3 Analysis on Individual Queries

The previous section reported that incorporating the trigger model into a reranking scheme significantly improved MAP by 22.0%. While the overall effect is evident, it does not tell how the trigger model affected the result for each topic. For more detailed analysis, Figure 4 presents average precision (AP) improvement after reranking using the trigger model. Parameter β was fixed to the optimum (0.35) identified above.

Examining the results, notable increase (> 0.1 points in AP) was observed for **macbook pro** (+0.2227), **mardi gras** (+0.1141), **heineken** (+0.1977), **shimano** (+0.1929), **zyrtec** (+0.1211), and **board chess** (+0.1333). On the other hand, there is a slight drop

[1]It does not mean that INM is superior to VSM. In fact, the best reported result by Zhang et al. [31] was obtained by a vector space model.

[2]This baseline was provided to TREC 2008 Blog Track participants to facilitate fairer comparison between participants and will be available to public after the conference is over. The specifications of the baseline have not been disclosed at the time of writing this paper.

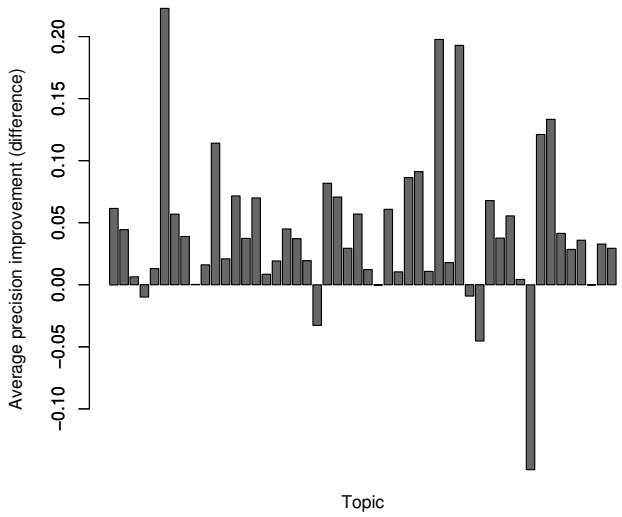

Figure 4: Average precision improvement after reranking for individual topics.

for **ann coulter** (−0.0099), **cindy sheehan** (−0.0327), **sonic food industry** (−0.0136), **west wing** (−0.0091), **world trade organization** (−0.0453), and **business intelligence resources** (−0.0002). Lastly, a large drop was found for **jim moran** (−0.1490). To help identify the commonalities, if any, underlying respective topic sets, the following summaries their descriptions, mainly from Wikipedia.[3] 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.
- **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.
- **Heineken** (#883) is a Dutch 5% abv pale lager made by Heineken International since 1868.
- **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.
- **Board chess** (#894) is the traditional game of chess using 32 pieces and played on a board having 64 black and white squares.

The above topics can be categorized as products except for Mardi Gras and board chess. 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 and beer, the model 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.

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

Next, regarding the topics for which system performance decreased:

- **Ann Coulter** (#854) is an American conservative political commentator, syndicated columnist, and best-selling author.
- **Cindy Sheehan** (#871) is an American anti-war activist, whose son, Casey, was killed during his service in the Iraq War on April 4, 2004.
- **West Wing** (#886) is an American television serial drama [. . .]. The series is set in the West Wing of the White House, the location of the Oval Office and offices of presidential senior staff [. . .].
- **Jim Moran** (#892) has represented the 8th congressional district of Virginia since 1991. He is a member of the Democratic Party.
- **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.
- **Business intelligence resources** (#898) are not a specific named entity but refer to any resources available for business intelligence.

Judging from these descriptions, the first four can be categorized as "politics" and the next two as "organizations." These categories of topics appear to be difficult to improve by using our language model, although the negative impact was relatively small except for "Jim Moran." Finally, the last topic (#898) is originally vague as indicated by the lowest average precision of 0.0008 by initial search and thus not well suited for assessing the effectiveness of our language model.

Taken altogether, these results imply 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 emphasized, however, that there are also many topics in these categories which were improved by applying our model, such as Abramoff Bush (+3.6%), Colbert Report (+12.1%), Muhammad cartoon (+25.9%), Bruce Bartlett (+29.3%), Qualcomm (+5.7%), Ariel Sharon (+43.9%), and McDonalds (+25.9%).

The next section applies the model adaptation technique introduced in Section 2.4 to examine if further/any improvement is achieved especially for the above difficult topics.

3.3.4 Model Adaptation

Based on the steps described in Section 2.4, we conducted additional experiments for opinion retrieval using model adaptation. The value of k was experimentally set to 50. Table 5 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

Table 5: Comparison of initial search result and those after reranking by the subjective trigger model with/without adaptation

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%

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 6 presents some of the most influential (newly identified) trigger pairs that were actually used for estimating adapted $P_E(\cdot)$.

Table 6: 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)

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.

Then, we looked at the "difficult" individual topics discussed in Section 3.3.3 to see if there was positive effect of newly identified trigger pairs on them. Table 7 shows their average precision scores and percent improvement as compared with those by the original (not 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. We will continue to study better use of trigger pairs for opinion retrieval.

Table 7: 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%

4. 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 is a relatively new theme of study partly motivated by the TREC Blog Track [18] introduced in 2006. In the track, opinion retrieval was tackled as one of the open task challenges. Other tasks include opinion polarity analysis and feed search [4]. For opinion retrieval, 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.

The first type of approaches are lexicon-based, automatically or manually constructing a subjective word/phrase list and use it for estimating the opinionatedness of a given blog post [16, 17, 27, 29]. For example, Hannah et al. [6] created a English word list from various linguistic sources and computed for each word the opinionated discriminability based on the relevance judgment from the TREC 2006 Blog Track opinion retrieval task. The weighted word list was then used as a query to measure the opinionated nature of each document. Their approach yielded the highest improvement of 15.87% in MAP over their initial search at the TREC 2007 Blog Track. As compared with this type of approaches, an advantage of our proposed approach is that our approach does not require any relevance judgment data but only a corpus of opinions, which is abundant. In the category of lexicon-based approaches, some research groups also considered the proximity of those words and/or first and second personal pronouns to query terms under consideration [26, 32], which in part motivated the present study.

The second type of approaches used supervised classifiers to identify opinions. To our knowledge, the most successful results were reported by Zhang et al. [30, 31], who collected a large number of opinionated and non-opinionated documents from the web, specifically `retails.com` 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 data set) is 0.2726. Although the result appears to outperform those reported in the present paper, our proposed approach, combined with a stronger baseline, could yield an even higher MAP of 0.3221 as discussed in Section 3.3.2. Because an initial search result strongly correlates to the performance of opinion retrieval [11, 12], one needs to make sure that the same baseline (initial search result) is utilized in comparing two different approaches to opinion retrieval.

Comparing with the existing work, 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.

5. 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 Blog06 test collection showed that by incorporating subjective trigger pairs, system performance increased by 22.0% in MAP. 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 (or the corpus used to build the model). To deal with it, we proposed a framework to dynamically update the trigger model to a given topic, which overall had positive effects for most topic types.

For future work, we plan to examine our proposed model in comparison with other approaches on the same initial search results. Also, we will explore alternative textual resources, including the Blog06 test collection and a larger set of Amazon reviews, for language modeling. Finally, we would like to investigate 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-based approach [22] taking words around the topic may be beneficial.

6. REFERENCES

[1] E. Adar and L. Adamic. Tracking information epidemics in blogspace. In *Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence*, pages 207–214, 2005.

[2] N. Agarwal, H. Liu, L. Tang, and P. S. Yu. Identifying the influential bloggers in a community. In *Proceedings of the international conference on Web search and web data mining*, pages 207–218, 2008.

[3] X. Ding, B. Liu, and P. S. Yu. A holistic lexicon-based approach to opinion mining. In *Proceedings of the international conference on Web search and web data mining*, pages 231–240, 2008.

[4] J. L. Elsas, J. Arguello, J. Callan, and J. G. Carbonell. 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*, pages 347–354, 2008.

[5] A. Esuli and F. Sebastiani. PageRanking WordNet synsets: An application to opinion mining. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, 2007.

[6] D. Hannah, C. Macdonald, J. Peng, B. He, and I. Ounis. University of Glasgow at TREC 2007: Experiments in blog and enterprise tracks with Terrier. In *Proceesings of the 16th Text Retrieval Conference*, 2007.

[7] N. Jindal and B. Liu. Opinion spam and analysis. In *Proceedings of the international conference on Web search and web data mining*, pages 219–230, 2008.

[8] D. Jurafsky and J. H. Martin. *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition*. Prentice Hall, 2000.

[9] R. Lau, R. Rosenfeld, and S. Roukos. Trigger-based language models: a maximum entropy approach. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 45–48, 1993.

[10] V. Lavrenko and B. Croft. Relevance based language models. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 120–127, 2001.

[11] C. Macdonald, B. He, I. Ounis, and I. Soboroff. Limits of opinion-finding baseline systems. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 747–748, 2008.

[12] C. Macdonald, I. Ounis, and I. Soboroff. Overview of the TREC-2007 blog track. In *Proceesings of the 16th Text Retrieval Conference*, 2007.

[13] C. D. Manning and H. Schütze. *Foundations of statistical natural language processing*. MIT Press, 1999.

[14] Q. Mei, X. Ling, M. Wondra, H. Su, and C. Zhai. Topic sentiment mixture: Modeling facets and opinions in weblogs. In *Proceedings of the 16th International World Wide Web Conference*, 2007.

[15] D. Metzler and W. Croft. 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, 2004.

[16] G. Mishne. Multiple ranking strategies for opinion retrieval in blogs. In *Proceesings of the 15th Text Retrieval Conference*, 2006.

[17] D. Oard, T. Elsayed, J. Wang, Y. Wu, P. Zhang, E. Abels, J. Lin, and D. Soergel. TREC-2006 at Maryland: Blog, enterprise, legal and QA tracks. In *Proceesings of the 15th Text Retrieval Conference*, 2006.

[18] I. Ounis, M. de Rijke, C. Macdonald, G. Mishne, and I. Soboroff. Overview of the TREC-2006 blog track. In *Proceesings of the 15th Text Retrieval Conference*, 2006.

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

[20] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., 1983.

[21] K. Seki, Y. Kino, S. Sato, and K. Uehara. TREC 2007 blog track experiments at Kobe University. In *Proceesings of the 16th Text Retrieval Conference*, 2007.

[22] K. Seki and J. Mostafa. An application of text categorization methods to gene ontology annotation. In *Proceedings of the 28th annual international ACM SIGIR conference on research and development in information retrieval*, pages 138–145, 2005.

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

[24] T. Tao and C. Zhai. 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*, pages 162–169, 2006.

[25] C. Tillmann and H. Ney. *Grammatical Interference: Learning Syntax from Sentences*, chapter Selection criteria for word trigger pairs in language modeling, pages 95–106. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, 1996.

[26] O. Vechtomova. Using subjective adjectives in opinion retrieval from blogs. In *Proceesings of the 16th Text Retrieval Conference*, 2007.

[27] K. Yang, N. Yu, A. Valerio, and H. Zhang. WIDIT in trec-2006 blog track. In *Proceesings of the 15th Text Retrieval Conference*, 2006.

[28] K. Yang, N. Yu, and H. Zhang. WIDIT in TREC 2007 blog track: Combining lexicon-based methods to detect opinionated blogs. In *Proceesings of the 16th Text Retrieval*

Conference, 2007.

[29] M. Zhang and X. Ye. 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*, pages 411–418, 2008.

[30] W. Zhang and C. Yu. UIC at TREC 2006 blog track. In *Proceesings of the 15th Text Retrieval Conference*, 2006.

[31] W. Zhang, C. Yu, and W. Meng. Opinion retrieval from blogs. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 831–840, 2007.

[32] G. Zhou, H. Joshi, and C. Bayrak. Topic categorization for relevancy and opinion detection. In *Proceesings of the 16th Text Retrieval Conference*, 2007.