Opinion Retrieval Experiments using Generative Models (Notebook Version)

Koji Eguchi* and Chirag Shah†
National Institute of Informatics
Tokyo 101-8430, Japan
{eguchi,chirag}@nii.ac.jp

Abstract

Ranking blog posts that express an opinion regarding a given topic should serve a critical function in helping users. We carried out experiments using three types of opinion retrieval methods in the framework of probabilistic language models. The first method combines topic-relevance model and opinion-relevance model that captures topic dependence of the opinion expressions. The second method makes use of probability that any of opinion-bearing words appear in each target document as document prior probability in query-likelihood model. The third method makes use of probability that any of adjectives or adverbs appear in each target document as document prior probability in the query-likelihood model, assuming opinionated documents tend to contain more adjectives or adverbs than other documents.

1 Introduction

The recent rapid expansion of access to information has significantly increased the demands on retrieval or classification of sentiment information from a large amount of textual data. The field of *sentiment classification* has recently received considerable attention, where the polarities of sentiment, such as positive or negative, were identified from unstructured text [11]. A number of studies have investigated sentiment classification at document level, e.g., [9, 2], and at sentence level, e.g., [4, 5, 8]; however, the accuracy is still less than desirable. Therefore, ranking according to the likelihood of containing sentiment information is expected to serve a crucial function in helping users.

For this objective, Eguchi and Lavrenko proposed *sentiment retrieval* models, aiming at finding sentences containing information with a specific sentiment polarity on a certain topic [3]. Intuitively, the expression of sentiment in text is dependent on the topic. For example, a negative view for some voting event may be expressed using 'flaw', while a negative view for some politician may be expressed using 'reckless'. Moreover, sentiment polarities are also dependent on topics or domains. For example, the adjective 'unpredictable' may have a negative orientation in an automotive review, in a phrase such as 'unpredictable steering', but it could have a positive orientation in a movie review, in a phrase such as 'unpredictable plot', as mentioned in [12] in the context of his sentiment word detection. The sentiment retrieval models are based on the framework of generative language modeling, not only assuming query terms expressing a certain topic, but also assuming that the polarity of sentiment interest is specified by the user in some manner, where the topic dependence of the sentiment is considered.

In [3], sentence level was focused in the experiments; however, the model can be applied to textual chunks of any length. For the TREC-2006 Blog Track, we modified the sentiment retrieval models for the opinion retrieval task. We also explored the use of some document features as document prior probability in query-likelihood model [10].

*Now also with Kobe University
†Now with University of North Carolina at Chapel Hill

2 A Generative Model of Opinion

2.1 Definitions

According to [3], we start by providing a set of definitions that will be used in the remainder of this section. The task of our model is to *generate* a collection of statements $\mathbf{w}_1 \ldots \mathbf{w}_n$. A statement $\mathbf{w}_i$ is a string of words $\mathbf{w}_{i1} \ldots \mathbf{w}_{in_i}$, drawn from a common vocabulary $\mathcal{V}$. We introduce a binary variable $b_{ij} \in \{S, T\}$ as an indicator of whether the word in the jth position of the ith statement will be a topic word or a opinion-bearing word. For our purposes, b_{ij} is determined heuristically (*automatic annotation*), in this paper.

As a matter of convenience we will often denote a statement as a pair $\{\mathbf{w}_i^s, \mathbf{w}_i^t\}$, where $\mathbf{w}_i^s$ contains the opinion-bearing words and $\mathbf{w}_i^t$ contains the topic words. As we mentioned above, the user's query is treated as just another statement. It will be denoted as a pair $\{\mathbf{q}^s, \mathbf{q}^t\}$, corresponding to opinion-bearing words and topic keywords. We will use $\mathbf{p}$ to denote a unigram language model, i.e., a function that assigns a number $\mathbf{p}(v) \in [0, 1]$ to every word v in our vocabulary $\mathcal{V}$, such that $\Sigma_v \mathbf{p}(v) = 1$. The set of all possible unigram language models is the probability simplex $I\!P$. We define $\pi : I\!P \times I\!P \rightarrow [0, 1]$ to be a measure function that assigns a probability $\pi(\mathbf{p}_1, \mathbf{p}_2)$ to a pair of language models $\mathbf{p}_1$ and $\mathbf{p}_2$.

2.2 Generative model

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

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

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

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

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

2.3 Using the model for retrieval

The generative model presented above can be applied to opinion retrieval in the following fashion. We start with a collection of statements C and a query $\{\mathbf{q}^s, \mathbf{q}^t\}$ supplied by the user. We use the procedure outlined in Section **2.2** to estimate the topic- and opinion-relevance models corresponding to the user's information need, and then determine which statements in our collection most closely correspond to these models of relevance. The topic-relevance model R_t and opinion-relevance model R_s are estimated in the similar fashion described in [3]. Once we have estimates for the topic and sentiment relevance models, we can rank testing statements $\mathbf{w}$ by their similarity to R_t and R_s. We rank statements using a variation of cross-entropy, which was proposed by [13]:

$$\alpha \sum_v R_t(v) \log \mathbf{p}_t(v) + (1-\alpha) \sum_v R_s(v) \log \mathbf{p}_s(v). \quad (2)$$

Here the summations extend over all words v in the vocabulary. A weighting parameter α allows us to change the balance of topic and sentiment in the final ranking formula; its value can be selected empirically.

3 Opinion Retrieval Task

3.1 Using opinion-relevance models

We define a variation of the sentiment retrieval model [3]. As input, we used (1) a set of topic keywords $\mathbf{q}^t$ and (2) a set of opinion-bearing seed words $\mathbf{q}^s$. Since we did not have a training data set, all the model parameters were the same as used in [3]. These model parameters are not very appropriate for the opinion retrieval experiments in the Blog Track, as we describe later in this paper.

We detected opinion-bearing words using lists of words. We used sentiment word list contained

in *OpinionFinder* [1], which consists of 2230 positive and 3913 negative words. We extracted opinion-bearing expressions using the list of words above.

3.2 Other models

NII1: As a baseline, we carried out experiments using Indri [7]. Entire corpus with blog documents was indexed. The topics were used as queries and top 1000 documents were retrieved using query likelyhood approach on the Indri platform.

NII7: As another baseline, we used (topic-) relevance model [6], which was estimated using the (weighted) mixuture of each model of a certain number of top-ranked documents. We used the result of the baseline run of NII1, and re-ranked them usign the topic-relevance model.

NII6: This is the retrieval model as described in Section **2.3**. We used the result of the baseline run of NII1, and re-ranked them using this retrieval model.

NII5: We obtained a list of opinion-bearing words and used it to find out the document prior probability in the language modeling framework. This probability was calculated by finding the total number of opinion-bearing words in a document and dividing it by the total number of words in that document. This probability was multiplied by the query likelihood probability. The query likelihood probability was obtained from the baseline run of NII1.

NII3: We made use of probability of any of adjectives or adverbs in each target document as document prior probability in addition to the query-likelihood model on the Indri platform, assuming opinionated documents tend to contain more adjectives or adverbs than other documents.

4 Results and Discussions

According to the relevance judgment results, NII5 and NII3 did not work, unfortunately. After our official runs were submitted we discovered some bugs in our implementation, though. As for NII7 and NII6, we used the model parameters estimated in [3], where sentence-level retrieval experiments were performed, because we could not use training data to estimate the model parameters. This setting was not appropriate for blog-post retrieval, and so the performance of NII7 and NII6 was not as good as that of NII1. Using the relevance judgment data given by the organizers, we are planning to estimate the model parameters and to perform the additional experiments to investigate how the topic-sentiment relevance model actually works at the appropriate setting.

Table 1: Mean average precision of our official runs

RunID	opinion-relevance	topic-relevance
NII1	0.0466	0.0834
NII7	0.0383	0.0736
NII6	0.0324	0.0645
NII5	0.0195	0.0475
NII3	0.0168	0.0419

Acknowledgments

This work was supported in part by the Overseas Research Scholars Program and the Grant-in-Aid for Scientific Research (#17680011 and #18650057) from the Ministry of Education, Culture, Sports, Science and Technology, Japan, Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect those of the sponsor.

References

[1] MPQA releases – corpus and opinion recogntion system. http://www.cs.pitt.edu/mpqa/.

[2] K. Dave, S. Lawrence, and D. M. Pennock. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of WWW-03, 12th International Conference on the World Wide Web*, pages 519–528, Budapest, HU, 2003. ACM Press.

[3] K. Eguchi and V. Lavrenko. Sentiment retrieval using generative models. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006)*, Jul. 2006.

[4] M. Hu and B. Liu. Mining and summarizing customer reviews. In *Proceedings of KDD '04, the ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177, Seattle, US, 2004. ACM Press.

[5] S.-M. Kim and E. Hovy. Determining the sentiment of opinions. In *Proceedings COLING-04, the Conference on Computational Linguistics*, Geneva, CH, 2004.

[6] V. Lavrenko and W. B. Croft. Relevance based language models. In *Proceedings of the 24th Annual International ACM SIGIR Conference*, pages 120–127, New Orleans, Louisiana, USA, Sep. 2001.

[7] D. Metzler and W. B. Croft. Combining the language model and inference network approaches to retrieval. *Information Processing and Management*, 40(5):735–750, 2004.

[8] K. Nigam and M. Hurst. *Computing Attitude and Affect in Text: Theory and Applications*, chapter Towards a Robust Metric of Opinion. Springer, 2005.

[9] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP-2002)*, pages 79–86, Philadelphia, Pennsylvania, USA, Jul. 2002.

[10] J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference*, pages 275–281, Melbourne, Australia, Aug. 1998.

[11] J. Shanahan, Y. Qu, and J. Wiebe, editors. *Computing attitude and affect in text*. Springer, 2005.

[12] P. D. Turney. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002)*, pages 417–424, Philadelphia, Pennsylvania, USA, Jul. 2002.

[13] C. Zhai. *Risk Minimization and Language Modeling in Text Retrieval*. PhD dissertation, Carnegie Mellon University, Pittsburgh, PA, July 2002.