The Robert Gordon University at the Opinion Retrieval Task of the 2007 Trec Blog Track

Rahman Mukras, Nirmalie Wiratunga, Robert Lothian

School of Computing
The Robert Gordon University
St Andrew Street
Aberdeen UK, AB25 1HG
{ram,nw,rml}@comp.rgu.ac.uk

Abstract. The Robert Gordon University (RGU) participated in the Opinion Retrieval Task of the Trec 2007 Blog Track. At the core of the system we developed is a set of training documents labeled with respect to opinion. These documents are used to train a classifier in order to classify the documents that are relevant to the given Trec topics. However, a major limitation with these training documents is that they are not always generic enough to serve as good training examples for all the 50 Trec topics. We therefore propose a solution that generalizes their content by exploiting the context of opinion related language constructs such as adjectives, verbs, and adverbs. Our system significantly improves over RGU's previous Blog Track systems.

1 Introduction

The simplicity of Internet publishing has resulted in individuals postings up their thoughts in a variety of different forms in an almost uncontrollable manner. Furthermore, many of these postings remain largely unmonitored. One particular publishing form of interest is Blogs (short for Web-log). Blogs have been claimed to be the latest form of self expression and it has been noted that they are rapidly changing the face of the Web.

The key aspect of Blogs that makes them attractive is that they are mainly authored by independent individuals, with the sole purpose of making their opinions known to the world. Consequently, Blogs are highly rich in opinion and are often in sync with current affairs. This factor makes them quite useful to industries such as marketing or the media where direct feedback is an invaluable resource.

The collection used in both the 2006 and 2007 Blog Tracks [4] consists of over a three million blog posts collected over 77 days. It was meant to be a realistic snapshot of the blogosphere and hence offers an excellent test-bed for Blog analysis research. This study addresses the opinion retrieval task of the 2007 Blog Track [5]. This task was first introduced in Trec 2006. It basically involves retrieving opinionated documents that are relevant to each of the 50 predefined Trec topics regardless of their opinion orientation. Each retrieved document should, however, be assigned to a real-valued *opinion score* in the range of $[0 \dots 1]$, where 0 signifies a neutral opinion, whereas 1 signifies an extreme opinion that could be either positive or negative.

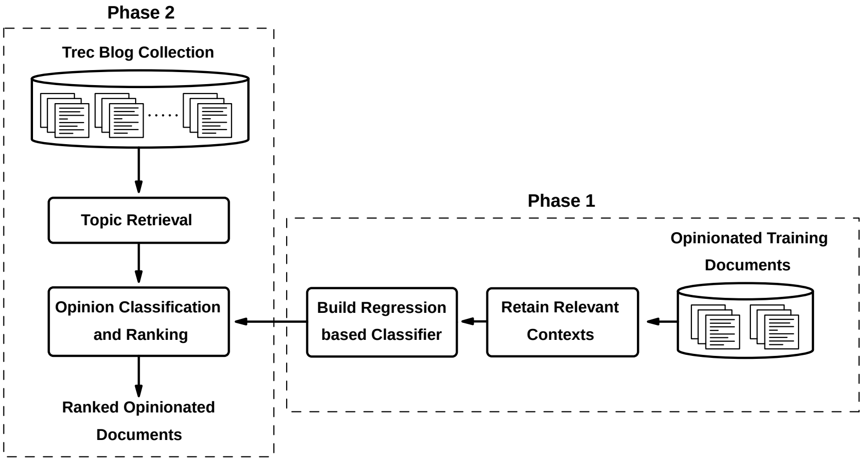
Phase 2
Trec Blog Collection
Topic Retrieval
Opinion Classification and Ranking
Ranked Opinionated Documents
Phase 1
Opinionated Training Documents
Build Regression based Classifier
Retain Relevant Contexts

Fig. 1. An Overview of the Opinion Retrieval System.

The idea we present to address this task is composed of two phases that draw from both Natural Language Processing (NLP) and Information Retrieval (IR). Our technique basically exploits the context of opinion related language constructs, such as adjectives verbs and adverbs, to identify and rank opinionated texts within the collection. These language constructs were chosen primarily because they are commonly associated with opinion rich contexts. For instance, an adjective such as "great" would typically occur within contexts such as "great player," or "great disaster."

In the first phase of the procedure, Part-of-Speech (PoS) tags are assigned to the words (features) contained in a set of training documents that are labeled with respect to opinion. These documents are then pruned such that only a set of selected words, along with their respective contexts, are retained. Proper nouns are also omitted as they tend to be domain specific [9]. This results in a corpus with a high precision of opinion rich phrases that are relatively domain independent. The documents of the resultant corpus are then mapped onto a vector space [8] after which a regression based Support Vector Machine (SVM) [3] is trained on the resultant document vectors.

In the second and last phase, a Lucene search engine is used to retrieve all documents, from the Trec Blog collection, that are relevant to the current topic of interest. The regression based SVM is then used to assign each of these documents to an opinion score, and this completes a single Trec run.

The remainder of this paper is organized as follows. Section 2 describes the two phase procedure in more detail. The Implementation specifications are then presented in Section 3. Finally, the Conclusions and Future Work appear in Section 4.

2 The Opinion Retrieval System

A complete overview of our opinion retrieval system has been illustrated in Fig. 1. A crucial resource for this system is the background set of training documents that are labeled with respect to opinion. Each of these labels assumes a value in the ordered set $\{c_1, \ldots, c_n\}$, where $c_1 < c_n$ and c_1, c_n respectively represent an extreme negative and positive opinion.

These training documents offer a good estimate of the structure and content of opinion rich texts. However, one limitation with them is that they may not be general enough to suffice as good document examples for each of the 50 Trec topics. To tackle this problem, we applied a procedure that retains the relevant contexts of language constructs such as adjectives verbs and adverbs. The hypothesis behind this is based on the fact that these constructs act as the transmitters, rather than the objects, of an opinion. Consequently, they would be used in much the same way across various domains. This type of pruning would therefore generalize the opinions expressed within the training documents.

To retain the relevant contexts, we first applied PoS tags to the text within the training documents by using the RASP PoS tagger [1]. We then retained a context of 10 words on either side of each word that was tagged as an adjective, verb, or adverb i.e: JJT, JJ, JJR, VV0, RR, RG, RGA, RGR[1]. We, however, did not retain any singular or plural proper nouns (NP, NP1, NP2) as these tend to be domain specific [9]. Once the training documents were pruned, we then mapped them onto a vector space whose dimensions, or features, were determined by the Information Gain measure [10]. A regression based SVM was then trained on the resultant document vectors.

Finally, given a list of documents from the Trec Blog Corpus that are relevant to a Trec topic, the regression based SVM was used to assign the i^{th} document in this list to a score q_i that assumes a real-value in the range $[c_1 \ldots c_n]$. Note, however, that the Trec rules require that the score q_i be mapped onto the range of $[0 \ldots 1]$, where 0 signifies a neutral opinion, whereas 1 signifies an extreme opinion that could be either positive or negative. In order to accomplish this, we mapped the score of the i^{th} document to the value $\left|2\left(\frac{q_i - c_1}{c_n - c_1}\right) - 1\right|$ which satisfies the Trec requirement.

3 Implementation

In order to prepare the Trec Blog collection, we first extracted the text from the initial HTML format discarding all tokens that contained non-printable characters. We then preprocessed it using the sequence of tokenization, conversion to lowercase, stemming and stopword removal (The 50 Trec topics also went through the same preprocessing steps). We finally indexed the resultant collection, containing 2,466,650 documents, using the Lucene[2] search engine. This entire procedure lasted 46 continuous days on standard hardware running on a Ubuntu Linux platform.

The second task was to prepare the opinionated training datasets. These were four in number namely: The Edmunds dataset [6] with classes $\{1, \ldots, 26\}$, the Rateitall dataset [2] with classes $\{1, \ldots, 5\}$, the Scale dataset [7] with classes $\{1, \ldots, 8\}$, and the documents that constituted the results of the Trec 2006 polarity task which had

[1] See the CLAWS2 Tagset: http://www.comp.lancs.ac.uk/ucrel/claws2tags.html

[2] http://lucene.apache.org

classes $\{0, \ldots, 4\}$. Note, however, that we only used classes 2, 3, and 4 which respectively correspond to a negative, a neutral, and a positive opinion. All four datasets were preprocessed in the similar fashion using the sequence of tokenization, PoS tagging, conversion to lower case, stemming, and stopword removal.

Once the four training datasets were ready, their contexts were pruned as discussed in the previous section. They were then used, in succession, to train a regression based SVM in order to classify the documents that were relevant to the 50 Trec topics. The outcome of the four train-classify sessions formed the basis of four of the runs that we submitted to Trec. The fifth run was based on plain relevance retrieval. The following list is a summary of all the five runs that we submitted:

1. rgu0: All opinion finding features turned off. Simply a Relevance run.
2. rgu1: Edmunds dataset used as background training data.
3. rgu2: Rateitall dataset used as background training data.
4. rgu3: Scale dataset used as background training data.
5. rgu4: Trec Polarity dataset used as background training data.

Although our official Trec results were far from being the best, our highest Mean Average Precision (MAP) of 0.2798 improved significantly from our previous years result of 0.0001. Hopefully this trend will continue for successive Trec competitions.

4 Conclusions and Future Work

In the Trec 2007 Block Track, our best run achieved a MAP of 0.2798. It also took seventeenth position among all the runs that were submitted by the 20 participants in terms of MAP, R-precision, b-Bref, and P@10. Although this performance leaves a great deal to be desired, our approach of exploiting the context of adjectives, verbs, and adverbs to identify opinionated text was quite innovative. For future work, we intend to build upon this approach by investigating the effect of variable sized contexts. We also intend to employ deeper NLP techniques to determine the focal point of a context, rather than simply assuming that it would always be an adjective, a verb, or an adverb.

References

1. Ted Briscoe and John Carroll. Robust Accurate Statistical Annotation of General Text. In *Proc. of LREC*, pages 1499–1504, Las Palmas, Canary Islands, May 2002.
2. Sutanu Chakraborti, Rahman Mukras, Robert Lothian, Nirmalie Wiratunga, Stuart Watt, and David Harper. Supervised Latent Semantic Indexing using Adaptive Sprinkling. In *Proc. of IJCAI*, pages 1582–1587. AAAI Press, 2007.
3. Thorsten Joachims. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In *Proc. of ECML*, pages 137–142, London, UK, 1998. Springer-Verlag.
4. Craig Macdonald and Iadh Ounis. The TREC Blogs06 Collection : Creating and Analysing a Blog Test Collection. Technical report, Department of Computing Science, University of Glasgow, Glasgow, UK, 2006.
5. Craig Macdonald, Iadh Ounis, and Ian Soboroff. Overview of the TREC-2007 Blog Track. In *Proc. of the TREC*, 2007.

6. Rahman Mukras, Nirmalie Wiratunga, Robert Lothian, Sutanu Chakraborti, and David Harper. Information Gain Feature Selection for Ordinal Text Classification using Probability Re-distribution. In *Proc. of IJCAI Textlink Workshop*, 2007.
7. Bo Pang and Lillian Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proc. of ACL*, pages 115–124, 2005.
8. G. Salton, A. Wong, and C. S. Yang. A Vector Space Model for Automatic Indexing. *Commun. ACM*, 18(11):613–620, 1975.
9. Peter D. Turney. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In *Proc. of ACL*, pages 417–424, Morristown, NJ, USA, 2002. ACL.
10. Yiming Yang and Jan O. Pedersen. A Comparative Study on Feature Selection in Text Categorization. In *Proc. of ICML*, pages 412–420. Morgan Kaufmann, 1997.