Using Passage-Based Language Model For Opinion Detection in Blogs

Malik Muhammad Saad Missen
Université de Toulouse
IRIT UMR 5505 CNRS
Toulouse, France
+ 33 5 61 55 72 65
Malik.Missen@irit.fr

Mohand Boughanem
Université de Toulouse
IRIT UMR 5505 CNRS
Toulouse, France
+ 33 5 61 55 74 16
Mohand.Boughanem@irit.fr

Guillaume Cabanac
Université de Toulouse
IRIT UMR 5505 CNRS
Toulouse, France
+ 33 5 61 55 72 73
Guillaume.Cabanac@irit.fr

ABSTRACT
In this work, we evaluate the importance of Passages in blogs especially when we are dealing with the task of Opinion Detection. We argue that passages are basic building blocks of blogs. Therefore, we use Passage-Based Language Modeling approach as our approach for Opinion Finding in Blogs. Our decision to use Language Modeling (LM) in this work is totally based on the performance LM has given in various Opinion Detection Approaches. In addition to this, we propose a novel method for bi-dimensional Query Expansion with relevant and opinionated terms using Wikipedia and Relevance-Feedback mechanism respectively. We also compare the impacts of two different query terms weighting (and ranking) approaches on final results. Besides all this, we also compare the performance of three Passage-based document ranking functions (Linear, Avg, Max). For evaluation purposes, we use the data collection of TREC Blog06 with 50 topics of TREC 2006 over TREC provided best baseline with opinion finding MAP of 0.3022. Our approach gives a MAP improvement of almost 9.29% over best TREC provided baseline (baseline4).

Categories and Subject Descriptors
H.3.3 **[Information Storage and Retrieval]:** Information Search and Retrieval - *Query formulation, Relevance feedback, Search process.*

General Terms
Experimentation

Keywords
Opinion Detection, Blogs, Language Modeling, Semantic Relatedness, Passages.

1. INTRODUCTION
Blogs are very important source of opinions which motivates researchers to focus on opinion detection in blogs. This is the reason TREC introduced a Blog track in 2006 known as TREC Blog Track with the release of blog data collection. The data collection is *148GB* in size [1]. TREC has also provided the query relevance assessments (*qrels*) for this collection. In our work we are using this data collection using topics of TREC2006. We are dealing with two TREC tasks in this work i.e. Baseline adhoc retrieval task and Opinion Finding task. However the focus of our work is on the latter. We propose a passage-based LM approach for retrieving opinion documents. This work is impressed by the previous work on passage-level LM [2, 3].

SAC'10, March 22-26, 2010, Sierre, Switzerland.

2. OUR APPROACH
As previously proposed in the literature, our approach can also be realized in three basic stages i.e. Data Pre-processing, Topic Retrieval and Opinion Detection stage. In Data Pre-processing phase, data collection is cleaned from unnecessary noisy data like removal of unnecessary HTML tags (like script, style, etc). Links are also removed in our case. For Topic-Relevance, we are using TREC provided strongest baseline and Opinion Finding approach is explained below.

2.1 Opinion-Finding
Our Opinion Detection approach consists of the following three sub-stages:

2.1.1 Query Expansion
We expand the query twice: once with hyperlinked *Proper Nouns* and *Named Entities* in Wikipedia page of the title of the query for set of relevant terms $\{L_{REL}\}$ and then with opinion terms $\{L_{OPIN}\}$ using a kind of relevance (opinion feedback in fact) feedback for top 10 opinionated and non-opinionated documents. Opinion terms are extracted from the relevant passages (selected using terms in $\{L_{REL}\}$) of these marked documents found around the relevant terms. Before this passage selection, all noisy passages are removed. Duplicates are removed from $\{L_{OPIN}\}$ and *Subjectivity* score of each opinion term is calculated using lexical resource *SentiWordNet* [4]**.** All terms from $\{L_{OPIN}\}$ with subjectivity score of zero (or if they do not exist in *SWN*) are removed from the set.

$$Subj(t) = \frac{NegSWN(t) + PosSWN(t)}{|t_{sense}|} \qquad (1)$$

In equation 1, $NegSWN(t)$ is the negative score of the term t (for all its senses), $PosSWN(t)$ is the positive score of term t (for all its senses) $|t_{sense}|$ is the total number of senses for term t.

Similarly *collection frequency (cf), passage frequency (pf)* and *document frequency (df)* of all the terms ($\{L_{OPIN}\}$ and $\{L_{REL}\}$) are calculated. Opinion scores are calculated in two different ways: like in eq. 2 (labelled *ALL* in results) and eq. 3 (labelled *FREQ* in results). Final scores of opinion terms and relevant terms are calculated as:

$$Opin(t) = (cf * pf * df) * Subj(t) \quad \textbf{(2)}$$

$$Opin(t) = (cf * pf * df) \ if \ Subj(t) >= 0.5$$
$$Opin(t) = Term \text{ is dropped } if \ Subj(t) < 0.5 \quad \textbf{(3)}$$

$$Score_{\text{Opin}}(t) = Opin(t)_{\text{Rel + Opin}} - Opin(t)_{\text{NonRel}} \quad \textbf{(4)}$$

$$\text{Rel}(t) = (cf * pf * df) \quad \textbf{(5)}$$

$$Score_{\text{Rel}}(t) = \text{Rel}(t)_{\text{Rel + Opin}} - \text{Rel}(t)_{\text{NonRel}} \quad \textbf{(6)}$$

All terms are ranked according to their final scores, filtered for most common terms. Top *30* opinion terms are selected for each topic in addition to their WordNet synonyms and manually prepared expansions & contractions. A term is given a final relevance score of zero if answer of eq. 6 results in a negative value. In the end, we merge both sets of query terms i.e. $\{L_{REL}\}$ and $\{L_{OPIN}\}$ to form a final set of query terms.

2.1.2 Passage-Based Language Model

In our work, we use three passage-based documents scoring functions that are realized using a *Unigram Language Model* shown as below:

$$Score_{Avg}(d) = \frac{1}{|P|} \sum_{i=1}^{|P|} p(q|gi) \quad \textbf{(7)}$$

$$Score_{Max}(d) = \max_{gi \in d} p(q \mid gi) \quad \textbf{(8)}$$

$$Score_{Lin}(d) = \sum_{i=1}^{|P|} p(q|gi) \quad \textbf{(9)}$$

Where $Score_{Avg}(d)$ is the average of scores of all passages within a document d for a given query q, $Score_{Max}(d)$ is the score given to document d for a query q on behalf of one of its passages having maximum score, and $Score_{Lin}(d)$ is a linear addition of scores of all passages; $|P|$ is the total number of passages within the document d, g_i is the ith passage and $p(q|g)$ is the probability of generating query q from passage g which can also be written as shown below:

$$p(q|g) = p(t_1|g) * p(t_2|g) * ... * p(t_N|g) = \prod_{i=1}^{N} p(t_i|g) \quad \textbf{(10)}$$

Using above eq. leads to a sparse matrix. To avoid this situation, we use a mixture of three language models: the *Collection Model*, the *Document Model* and the *Passage Model* itself.

$$p(t_i|MIX) = \sum_{ti \in q} \lambda_1 p(t_i|g) + \lambda_2 p(t_i|d) + \lambda_3 p(t_i|C) \quad \textbf{(11)}$$

Where $\lambda_1 + \lambda_2 + \lambda_3 = 1$ and

$$p(t_i|g) = C(t_{ig}) * Score(t_i) / |T_g| \quad \textbf{(12)}$$

$$p(t_i|d) = C(t_{id}) * Score(t_i) / |T_d| \quad \textbf{(13)}$$

$$p(t_i|c) = C(t_{ic}) * Score(t_i) / |T_c| \quad \textbf{(14)}$$

Where $C(t_{iX})$= counts of term t_i in X, $|T_X|$=total terms in X. At the end, each document is given an opinion score which is later on added with the document relevance score (given in baseline) to give us a final score for the document. Finally the documents are re-ranked according to this final score.

3. RESULTS AND CONCLUSIONS

The results show an improvement of almost 9.29% (0.33) in *MAP* over baseline results. It's very clear that the results for ranking functions *Avg* and *Max* are far beyond from the results of *Linear* ranking function. It should be noted here that both functions i.e. *Avg* and *Max* are basically representing the score of one passage of a document while *Linear* ranking function is basically representing all the passages of a document which, in a way, proves our point that it's not the whole document which can improve the opinion retrieval but it may be only one passage that might be talking about the query. Our approach is giving promising results on TREC Blog 2006 collection but it would be better to evaluate it with more than one baseline on different data collections with few improvements in Query Expansion process and utilizing few other document or passage-based opinion evidences like document homogeneity.

Table 1. MAP and P@10

Ranking Function	Metric	ALL	FREQ
Avg	**MAP**	**0.3303**	0.2735
	P@10	0.6340	0.4980
Max	**MAP**	**0.3290**	0.2636
	P@10	**0.6340**	**0.5280**
Linear	**MAP**	0.2342	**0.2418**
	P@10	0.5160	0.5400

4. REFERENCES

[1] I. Ounis , C. Macdonald, and I. Soboroff, "Overview of the TREC-2008 NIST Blog Track", TREC 2008

[2] Michael Bendersky, "Passage Language Models in Adhoc Document Retrieval", Master's Research Thesis, Israel Institute of Technology Haifa, Israel, July 2007

[3] Xiaoyong Liu and W. Bruce Croft. Passage retrieval based on language models. 11th International Conference on Information and Knowledge Management (CIKM02)

[4] A. Esuli, and F. Sebastiani, "SentiWordNet: A publicly available lexical resourcefor opinion mining", LREC-06: in Proceedings of Language Resources and Evaluation Conference, European Language Resources Association, Genova, 2006