Integrating Proximity to Subjective Sentences for Blog Opinion Retrieval

Rodrygo L.T. Santos, Ben He, Craig Macdonald, and Iadh Ounis

Department of Computing Science
University of Glasgow, G12 8QQ, UK
{rodrygo,ben,craigm,ounis}@dcs.gla.ac.uk

Abstract. Opinion finding is a challenging retrieval task, where it has been shown that it is especially difficult to improve over a strongly performing topic-relevance baseline. In this paper, we propose a novel approach for opinion finding, which takes into account the proximity of query terms to subjective sentences in a document. We adapt two state-of-the-art opinion detection techniques to identify subjective sentences from the retrieved documents. Our first technique uses the OpinionFinder toolkit to classify the subjectiveness of sentences in a document. Our second technique uses an automatically generated dictionary of subjective terms derived from the document collection itself to identify the most subjective sentences in a document. We extend the Divergence From Randomness (DFR) proximity model to integrate the proximity of query terms to the subjective sentences identified by either of the proposed techniques. We evaluate these techniques on five different strong baselines across two different query datasets from the TREC Blog track. We show that we can significantly improve over the baselines and that, in several settings, our proposed techniques can at least match the top performing systems at the TREC Blog track.

1 Introduction

Blogs have recently emerged as a new open, rapidly evolving, and promptly reactive publishing medium on the World Wide Web [1]. The blogosphere – the collection of blogs on the Web – constitutes a fascinating ground for researchers in different areas and, in particular, for Information Retrieval. The information seeking behaviour in the blogosphere differs from that of traditional Web searchers [2]: blog searchers are mainly interested in uncovering the public sentiment towards named entities (e.g., people, organisations, locations) and in information on broad, conceptual topics (e.g., politics, sports, religion). In common, these information needs are of a subjective nature, characterising a distinctive and challenging feature of blog search.

Retrieving documents that are not only relevant to a given topic but that also contain an opinion about the topic has been shown to be a difficult task. Since 2006, the Text REtrieval Conference (TREC) has been running a Blog track and a corresponding opinion-finding task to address this problem [3,4].

M. Boughanem et al. (Eds.): ECIR 2009, LNCS 5478, pp. 325–336, 2009.

An important issue while evaluating the performance of different opinion retrieval systems is to look at how they perform over strong baselines that do not apply any specific opinion-finding technique. After the first two editions, however, the experiments conducted within the Blog track opinion-finding task have shown that outperforming a strong topic-relevance baseline remains a challenge. Indeed, while some participants were able to show a marked increase in performance when using opinion detection features on top of good topic-relevance baselines, other groups did not manage to improve their baselines [5]. In fact, a recent study [6] showed that some stronger topic-relevance baselines could not be improved even by applying the most effective opinion-finding approaches.

In order to address this problem, we adapt two opinion-finding techniques that were shown to perform well across different baselines. In particular, we propose an approach that integrates the proximity between query terms and subjective sentences – as identified by either of these two techniques – to existing retrieval systems in order to improve their opinion retrieval performance. We argue that this proximity can be seen as an estimate of the extent to which a given document is not only about the entity represented by the query, but also expresses an opinion towards it. We evaluate the effectiveness of our approach using the devised techniques for identifying subjective sentences in documents and show that they provide statistically significant improvements over five strong baselines across two different query datasets. In addition, we show that our proposed approach provides comparable opinion retrieval performances to the state-of-the-art approaches from the TREC 2008 Blog track opinion-finding task.

The remainder of this paper is organised as follows. In Section 2, we introduce the TREC paradigm for experimentation on opinion retrieval and survey the main related approaches to our work. In Section 3, we describe our two techniques to identify subjective sentences in text and introduce our proximity model for modifying the scores of documents based on the proximity of query terms to the subjective sentences identified in these documents. In Sections 4 and 5, we describe the experimental setup and the results of the evaluation of our approach over five strong baselines. Finally, in Section 6, we present our final remarks.

2 Blog Opinion Retrieval at TREC

The TREC Blog track opinion-finding task addresses a search scenario where a user aims to uncover what the bloggers are saying or thinking about a named entity X. In summary, the user's intention is to take the "pulse of the blogosphere" on a topic X. The task has been running in TREC since the Blog track inception in 2006. All experiments in the Blog track have been done using the Blogs06 collection, representing a large sample crawled from the blogosphere over an eleven week period from December 6, 2005 until February 21, 2006 [7]. The collection is 148GB in size, with three main components consisting of 38.6GB of XML feeds (i.e., the blog), 88.8GB of permalink documents (i.e., a single blog post and all its associated comments) and 28.8GB of HTML homepages (i.e., the main entry to the blog). In this paper, we follow the TREC setting [3,4] and experiment using the permalink documents as retrieval units. There are over 3.2 million permalink documents in the Blogs06 collection.

Each participating system is evaluated using a set of topics and their associated relevance assessments. An example topic is shown in Figure 1. The relevance assessment procedure for the documents retrieved for the topics has two levels. The first level assesses whether a given blog post, i.e., a permalink, contains information about the target and is therefore relevant. The second level assesses the opinionated nature of the blog post if it was deemed relevant in the first assessment level [3,4].

```
<top>

<num> Number: 1030 </num>
<title> System of a Down </title>

<desc> Description:
What do people think about the metal band System of a Down and
their music?
</desc>

<narr> Narrative:
Any positive or negative comment about System of a Down,
their music, albums, or songs is relevant. Opinions concerning
the performers' personal lives or endorsements are not relevant.
</narr>

</top>
```

Fig. 1. TREC 2008 Blog track opinion-finding task, topic 1030

One of the lessons learnt from the TREC 2006 and 2007 Blog tracks is that a good performance in opinion finding is strongly dominated by the underlying topic-relevance performance [3,4]. In fact, it has been shown that outperforming a strong topic-relevance baseline remains a challenge.

Indeed, after two years of the TREC Blog track, only a few opinion detection approaches have been shown to be effective in enhancing reasonably good topic-relevance baselines. Among the most effective approaches, the dictionary-based one consists in automatically building a weighted dictionary of opinionated terms derived from the target collection. The top weighted terms from the resulting dictionary are then submitted as a query to generate an opinionated score for each document of the collection [8]. A similar approach was used by Amati et al., who proposed a semi-automatic method for learning an opinion dictionary from the Blogs06 collection [9]. A different technique consists in using a pre-compiled list of subjective terms and computing their proximity to the query terms in each retrieved document of the collection [10,11]. Finally, another effective approach uses OpinionFinder, a freely available subjectivity analysis toolkit, to identify and score the opinionated nature of documents [12].

It was noted that most of the participating groups in TREC 2006 and 2007 Blog tracks approached the opinion-finding task as a re-ranking problem [3,4]. In the first stage, the retrieval system aims to find as many relevant documents as possible regardless of their opinionated nature, while in the second stage, the system re-ranks those documents using some opinion detection technique and an appropriate combination of scores. As a consequence, in order to draw a better understanding of the most effective and stable opinion-finding techniques, in TREC 2008, the organisers provided the participating groups with 5 standard baselines, each covering a set of 100 topics from the first two editions of the TREC Blog track opinion-finding task and 50 new topics from its 2008 edition [13].

In the next sections, we propose and evaluate a novel proximity approach to opinion finding that uses a more coarse-grained evidence. In particular, instead of estimating the proximity of query terms to opinionated terms as some of the aforementioned techniques attempted, we extend the OpinionFinder and the dictionary-based approaches to identify subjective sentences from the retrieved documents and boost the scores of these documents when and if the query terms occur in close proximity with the identified sentences.

3 Proximity to Subjective Sentences

In this section, we describe our approach to integrate the proximity of query terms to subjective sentences – identified from the documents in the first-pass retrieval – in order to improve the effectiveness of existing blog opinion retrieval systems. In order to show the generality of our approach, we employ two different opinion-finding techniques, each having a different trade-off between efficiency and sentence identification effectiveness. Indeed, our approach is general in that different techniques can be employed in order to identify a given sentence as being either objective or subjective. We describe these techniques, and evaluate their effectiveness in the next sections.

3.1 NLP-Based Subjectiveness Classification

Our first technique to identify subjective sentences in documents uses OpinionFinder [14], a subjectivity analysis system aimed at supporting Natural Language Processing (NLP) applications by providing them with information about opinions expressed in text and also who expresses them.

OpinionFinder operates as a two-stage pipeline. The first stage performs general-purpose document processing, including semantic and part-of-speech tagging, named entity identification, tokenisation, stemming, and sentence splitting. The text is then parsed again to generate dependency parse trees. Finally, subjective terms and expressions are identified based on a large dictionary.

The second stage is responsible for the subjectivity analysis itself. It employs a Naive Bayes classifier to distinguish between objective and subjective sentences. This classifier is trained on sentences automatically generated from a large corpus of unannotated data by two rule-based classifiers. The result of the subjectivity analysis is manifested in the form of markup tags added to the original documents. After the whole collection is parsed, we index it by considering the sentence tags generated by OpinionFinder as special position markers, so that we can record the positions of every index term with respect to the sentences in which it occurs within a given document. Moreover, the classification of each of these sentences as either objective or subjective and the confidence *sw* of such classification as reported by OpinionFinder can be integrated directly in our approach as discussed in Section 3.3.

3.2 Dictionary-Based Subjectiveness Estimation

Our second technique to identify subjective sentences in documents from first-pass retrieval relies on a light-weight, dictionary-based approach [8]. In this approach, a dictionary of subjective terms is automatically derived from the target collection without requiring any manual effort. First of all, from the list of all terms in the collection ranked by their within-collection frequency in descending order, a skewed query model is applied to filter out those that are too frequent or too rare [15]. This aims at removing terms with too little or too specific information and which cannot be interpreted as generalised, query-independent opinion indicators. Using a training set of queries, the remaining terms from the list are weighted based on the divergence of their distribution in the set of opinionated documents retrieved for these queries against that in the set of relevant documents retrieved for the same set of queries.

During retrieval time, for each document returned in response to a given query, we estimate an aggregated subjectiveness weight *sw* for each of its sentences according to the formula:

$$sw = \frac{1}{|s|} \times \sum_{t \in s} dtw \tag{1}$$

where $t \in s$ corresponds to the set of all terms t in sentence s, $|s|$ is the number of terms in s, and dtw corresponds to the weight of term t according to the generated dictionary of subjective terms.

Sentence subjectiveness weights are then normalised by the maximum weight among all sentences in the document. Finally, sentences with a weight greater than a predefined threshold δ – a free parameter – are considered as subjective sentences. It is worth noting that OpinionFinder implicitly uses a similar mechanism in order to decide whether a given sentence is subjective or not. Using either technique, our approach integrates the proximity of query terms to the identified subjective sentences from the retrieved documents in order to re-rank them, as discussed in the next section.

3.3 Modelling Proximity to Subjective Sentences

Previous studies have shown that taking into account the dependency of query terms in documents can improve retrieval performance [16,17,18]. Our approach applies this idea in a novel way, namely, by boosting the scores of the retrieved documents based on the proximity between the query terms and the subjective sentences identified in each of these documents. Besides the fact that documents containing subjective sentences are more likely to be opinionated, our intuition is that the proximity of the query terms to the subjective sentences in the document provides a further indication that the document expresses an opinion about the topic of the query.

Given a retrieved document d and a set of subjective sentences S_d identified in this document by either the OpinionFinder or the dictionary-based approach (or any equivalent one), the score of document d with respect to a query Q is boosted according to the following linear combination:

$$\mathit{score}(d,Q) = \lambda_1 \times \mathit{score}(d,Q) + \lambda_2 \times \sum_{t \in Q} \sum_{s \in S_d} \mathit{prox}(t,s) \tag{2}$$

where $\mathit{score}(d,Q)$ is the score of the document d retrieved against a query Q, $t \in Q$ corresponds to the set of all query terms, $s \in S_d$ is the set of all subjective sentences in document d, $\mathit{prox}(t,s)$ is the proximity score assigned to the query term t and the subjective sentence s in document d, and λ_1 and λ_2 are free parameters of the linear combination.

In Equation (2), the document score $\mathit{score}(d,Q)$ can be estimated by any weighting model. For example, we can use the DFR PL2 document weighting model [19]. To efficiently compute the proximity score $\mathit{prox}(t,s)$, we use the pBiL DFR model, which does not consider the collection frequency of the pair $\langle t,s \rangle$. It is based on the binomial randomness model [20] and is computed as follows:

$$\begin{aligned}\mathit{prox}(t,s) = \mathit{qtw} \times \mathit{sw} \times \frac{1}{\mathit{pf}+1} \times \Big[& -\log_2(\mathit{wc}+1) \\ & + \log_2(\mathit{pf}+1) \\ & + \log_2(\mathit{wc}-\mathit{pf}+1) \\ & - \mathit{pf} \times \log_2 \frac{1}{\mathit{wc}} \\ & - (\mathit{wc} \times \mathit{pf}) \times \log_2 \left(1 - \frac{1}{\mathit{wc}}\right) \Big]\end{aligned} \tag{3}$$

where qtw and sw are the weights of the query term t and the sentence s, respectively, $\mathit{wc} > 0$ is the number of windows of size ws sentences in document d – where ws is a free parameter – and pf is the frequency of the pair $\langle t,s \rangle$ within windows of size ws sentences in the document. The procedure for training the parameters in our approach is detailed in the next section.

4 Experimental Setup

In the evaluation of our proposed approach for blog opinion retrieval, we use the Terrier Information Retrieval platform for both indexing and retrieval [21]. The remainder of this section details our experimental setup and the training settings for our approach and its underlying techniques as described in Section 3. Our evaluation results are discussed in Section 5.

4.1 Collection and Topics

Our experiments are based on the Blogs06 collection [7], which is currently the only available blog test collection with relevance assessments. Following the official TREC Blog track opinion-finding task setting [3,4], in which permalinks – i.e., blog posts and their associated comments – are used as the retrieval units, we parse the permalinks component of the Blogs06 collection using OpinionFinder and index it using Terrier. Each token is stemmed using Porter's English stemmer and standard English stopwords are removed.

We use two realistic, chronologically organised query datasets in our experiments. In the first dataset, we use the 50 topics from the TREC 2006 Blog track opinion-finding task, numbered 851 to 900, for training, and the 50 topics from TREC 2007, numbered 901 to 950, for testing. In the second dataset, we use the 100 topics from both TREC 2006 and 2007 Blog track opinion-finding tasks for training and the 50 topics from TREC 2008, numbered 1001 to 1050, for testing. Each topic comprises three fields, namely, title, description, and narrative. We only use the title topic field, which is usually short and resembles real user queries in practice as well as the official TREC Blog track setting [3,4].

4.2 Retrieval Baselines

As described in Section 2, in 2008, TREC provided 5 strongly performing, yet statistically different baselines. Each of these baselines covers all 2006 through 2008 topics and comprises a list of relevant documents produced by a "black box" search engine that retrieves as many relevant documents as possible without applying any specific opinion-finding feature [13]. We apply the different techniques used by our approach over each of these 5 topic-relevance baselines.

4.3 Training Setting

We tuned the parameters used by our approach on the training topics of our query datasets (i.e., the 50 topics from the TREC 2006 opinion-finding task and the 100 topics from the TREC 2006 and 2007 tasks, respectively) over the 5 standard baselines. The parameters selected for training were the threshold δ, used by the dictionary-based technique to select subjective sentences in a given document, the weight λ_2, used for combining the original document score with its proximity score (see Equation (2)), and the window size ws, used while counting the number of windows in which the query terms co-occur with subjective sentences in a document. The combination weight λ_1 in Equation (2) was fixed to 1, since the optimal setting is only related to the ratio between λ_1 and λ_2.

For the ws parameter, after performing a sweeping over the values in the range 1 through 30 and observing no marked effect on the retrieval performance of our approach over the baselines, we fixed its value to 5, since this setting resulted in the best performances across all baselines.

Once the ws parameter was fixed, we optimised the weight λ_2 for our NLP-based technique using OpinionFinder over each of the five baselines by maximising opinion mean average precision (MAP) on the training topics for each query dataset. For the dictionary-based technique, we optimised both the λ_2 and δ parameters using the same procedure.

5 Experimental Results

In this section, we discuss the evaluation results of our approach on the two query datasets described in the previous section. Since the opinion-finding task is an adhoc-like retrieval task, the primary measure for evaluating the retrieval performance of the participating groups is the mean average precision (MAP). Other metrics reported for the opinion-finding task are R-Precision (R-prec), binary Preference (bPref), and Precision at 10 documents (P@10).

Using the TREC 2007 dataset, Table 1 summarises the retrieval performances of the two alternative techniques used by our approach over the 5 considered baselines in terms of topic relevance and opinion finding. Besides the standard TREC baselines, we compare the techniques used by our approach to their direct application for opinion retrieval as discussed in Section 2, i.e., without the use of proximity. In Table 1, OF and Dict stand for the OpinionFinder and dictionary-based approaches without the use of proximity while OFProx and DictProx stand for the addition of our sentence-level proximity approach to these techniques. The best value in each column for each baseline is underlined. A star (∗) indicates a significant improvement over the corresponding baseline according to the Wilcoxon signed-rank matched-pairs test at the 0.01 level, while a bullet (•) indicates a significant decrease in performance with respect to the baseline. Finally, a dagger (†) indicates a significant improvement of a technique over its

Table 1. Topic-relevance and opinion-finding performance results over 5 standard baselines for the TREC 2007 Blog opinion-finding task topics 901-950

	topic relevance				opinion finding			
	MAP	R-prec	bPref	P@10	MAP	R-prec	bPref	P@10
baseline1	0.4043	0.4305	0.4808	0.7620	0.2758	0.3226	0.3247	0.4540
+Dict	0.4202∗	0.4384	0.4917	0.7700	0.2988∗	0.3326	0.3406	0.5600
+DictProx	0.4307∗	0.4524∗	0.4934∗	0.8020	0.2990∗	0.3432∗	0.3366∗	0.5600
+OF	0.4117	0.4448	0.5025∗	0.7480	0.3120	0.3482	0.3644∗	0.5580
+OFProx	0.4610∗†	0.4759∗†	0.5287∗†	0.8140	0.3426∗†	0.3906∗†	0.3850∗†	0.6060
baseline2	0.3881	0.4113	0.4502	0.7220	0.3034	0.3461	0.3366	0.5320
+Dict	0.3886	0.4111	0.4522∗†	0.7200	0.3020	0.3474	0.3362•†	0.5540
+DictProx	0.3878•	0.4119•	0.4508∗	0.7300	0.3027•	0.3469∗	0.3359•	0.5380
+OF	0.3855	0.4168	0.4603∗	0.7120	0.3065	0.3549	0.3515∗	0.5400
+OFProx	0.4016∗†	0.4290∗†	0.4662∗†	0.7360	0.3274∗†	0.3624∗†	0.3579∗†	0.5820
baseline3	0.4619	0.4744	0.5066	0.7900	0.3489	0.3850	0.3705	0.5760
+Dict	0.4648∗	0.4754	0.5158∗†	0.7820	0.3561∗	0.3834	0.3757∗†	0.6140
+DictProx	0.4665∗	0.4795∗	0.5104∗	0.8040	0.3506∗	0.3785•	0.3638•	0.6020
+OF	0.4657	0.4821	0.5248∗	0.8060	0.3665	0.3983	0.3934∗	0.6260
+OFProx	0.4764∗†	0.4903∗	0.5249∗	0.8000	0.3703∗†	0.4006∗	0.3915∗	0.6220
baseline4	0.5303	0.5384	0.6483	0.8240	0.3784	0.4052	0.4403	0.5340
+Dict	0.5339	0.5462	0.6512	0.8160	0.3885	0.4116	0.4488	0.5700
+DictProx	0.5312∗	0.5381	0.6485∗	0.8240	0.3778•	0.4064	0.4397•	0.5340
+OF	0.5362∗	0.5470	0.6546∗	0.8380	0.3926∗	0.4214	0.4540∗	0.5680
+OFProx	0.5456∗†	0.5540∗	0.6542∗	0.8340	0.3968∗†	0.4190∗	0.4492∗	0.5720
baseline5	0.5465	0.5645	0.6507	0.8620	0.3805	0.4244	0.4364	0.5580
+Dict	0.5438	0.5686	0.6544	0.8800	0.3839	0.4220	0.4391	0.6280
+DictProx	0.5535∗	0.5717∗	0.6517	0.8720	0.3918∗	0.4253∗	0.4351	0.6140
+OF	0.5357	0.5588	0.6446	0.8600	0.3893	0.4254	0.4466	0.6240
+OFProx	0.5452	0.5679	0.6433	0.8520	0.4082	0.4451	0.4509	0.6500

Table 2. Topic-relevance and opinion-finding performance results over 5 standard baselines for the TREC 2008 Blog opinion-finding task topics 1001-1050

	topic relevance				opinion finding			
	MAP	R-prec	bPref	P@10	MAP	R-prec	bPref	P@10
baseline1	0.4032	0.4345	0.4408	0.7320	0.3239	0.3682	0.3514	0.5800
+Dict	0.4174∗	0.4460∗	0.4505∗	0.7440	0.3512∗†	0.3887∗	0.3797∗†	0.6380
+DictProx	0.4249∗†	0.4686∗†	0.4644∗†	0.7440	0.3497∗	0.3940∗†	0.3751∗	0.6320
+OF	0.4073∗	0.4392	0.4439	0.7460	0.3526∗	0.3954	0.3848	0.6460
+OFProx	0.4115∗	0.4492	0.4526∗	0.6880	0.3529∗	0.3919	0.3797∗	0.6040
baseline2	0.3107	0.3493	0.3411	0.6480	0.2639	0.3145	0.2902	0.5500
+Dict	0.3029	0.3414	0.3372	0.6400	0.2621	0.3087	0.2908	0.5660
+DictProx	0.3212∗†	0.3677∗†	0.3484∗†	0.6560	0.2758∗†	0.3266∗†	0.2984∗†	0.5540
+OF	0.3123	0.3536	0.3504	0.6360	0.2712	0.3225	0.3014	0.5540
+OFProx	0.3048	0.3474	0.3361	0.6080	0.2692	0.3151	0.2937	0.5420
baseline3	0.4343	0.4608	0.4662	0.6440	0.3564	0.3887	0.3677	0.5540
+Dict	0.4391∗	0.4615∗	0.4626•	0.7280∗	0.3669∗	0.3891∗	0.3728∗	0.6340∗
+DictProx	0.4379∗	0.4681∗	0.4659•	0.7160∗	0.3631∗	0.3949∗	0.3720∗	0.6180∗
+OF	0.4419∗	0.4682	0.4677∗	0.6980	0.3728∗	0.4017	0.3840∗	0.6060
+OFProx	0.4315	0.4625∗	0.4563	0.7000	0.3685	0.3965∗	0.3709	0.6200
baseline4	0.4724	0.4993	0.5127	0.7440	0.3822	0.4284	0.4112	0.6160
+Dict	0.4750	0.4962	0.5127	0.7520	0.3964	0.4370	0.4236	0.6400
+DictProx	0.4755	0.4976	0.5103	0.7640	0.3914	0.4361	0.4131	0.6300
+OF	0.4710	0.4899	0.5081	0.7640	0.3963	0.4370	0.4252	0.6600
+OFProx	0.4431	0.4730	0.4773	0.6940	0.3752	0.4134	0.3949	0.5980
baseline5	0.3745	0.4170	0.4342	0.7040	0.2988	0.3524	0.3395	0.5300
+Dict	0.3706•	0.4124•	0.4284•	0.6920	0.3008∗	0.3514•	0.3375•	0.5600
+DictProx	0.3832∗†	0.4264∗	0.4377∗	0.6880	0.3110∗†	0.3658∗	0.3463∗	0.5540
+OF	0.3777∗	0.4214∗	0.4363∗	0.7020	0.3098∗	0.3612∗	0.3472∗	0.5660
+OFProx	0.3894∗†	0.4359∗†	0.4409∗	0.7040	0.3312∗†	0.3829∗†	0.3603∗	0.6160

counterpart (i.e., OF vs. OFProx and Dict vs. DictProx). The performance of each of the 5 standard baselines is also presented.

From Table 1, we can see that the two techniques deployed by our approach significantly improve over the baselines in 8 out of 10 cases in terms of topic-relevance MAP, and in 7 out of 10 cases in terms of opinion MAP. In the latter case, it is also interesting to observe that the performances of our techniques over baseline3, baseline4, and baseline5 are superior than all but the best performing system at TREC 2007, which achieved an opinion MAP of 0.4341 [3].

Our approach based on OpinionFinder (OFProx) significantly improves over its counterpart (OF) in 4 out of the 5 baselines in terms of both topic-relevance and opinion MAP, while our dictionary-based approach (DictProx) was not significantly different from its base approach (Dict) across the considered baselines. Indeed, in further tests (not shown in Table 1), OFProx was significantly superior than DictProx over 3 baselines in terms of topic-relevance MAP and across all 5 baselines in terms of opinion MAP. On the other hand, there is no significant difference between the performance of their corresponding base approaches (OF and Dict) in terms of topic-relevance MAP across all baselines. In terms of opinion MAP, OF can only significantly outperform Dict in 2 out of the 5 baselines. This shows that our model was effective while using the subjective sentences identified by OpinionFinder, although it could not deliver the same benefit for the dictionary-based technique on this dataset.

In Table 2, we present similar results on our second dataset, the TREC 2008 Blog track opinion-finding task topics. Again, we show the retrieval performance of our approach using two techniques over the 5 previously described baselines.

Table 3. Opinion MAP over 5 standard baselines and average improvement for the TREC 2007 and 2008 Blog opinion-finding task topics

2007 topics	baseline1	baseline2	baseline3	baseline4	baseline5	improvement	
Baseline	0.2758	0.3034	0.3489	0.3784	0.3805	avg	stdev
+Dict	0.2988	0.3020	0.3561	0.3885	0.3839	2.70%	3.37%
+DictProx	0.2990	0.3027	0.3506	0.3778	0.3918	2.30%	3.66%
+OF	0.3120	0.3065	0.3665	0.3926	0.3893	5.05%	4.76%
+OFProx	0.3426	0.3274	0.3703	0.3968	0.4082	10.08%	7.99%
TREC median	0.3077	0.3298	0.3709	0.4128	0.3950	9.12%	4.07%
2008 topics	baseline1	baseline2	baseline3	baseline4	baseline5	improvement	
Baseline	0.3239	0.2639	0.3564	0.3822	0.2988	avg	stdev
+Dict	0.3512	0.2621	0.3669	0.3964	0.3008	3.02%	3.50%
+DictProx	0.3497	0.2758	0.3631	0.3914	0.3110	4.17%	2.39%
+OF	0.3526	0.2712	0.3728	0.3963	0.3098	4.72%	2.40%
+OFProx	0.3529	0.2692	0.3685	0.3752	0.3312	4.67%	5.18%
TREC median	0.3493	0.2705	0.3705	0.3846	0.3010	0.76%	0.73%

From Table 2, we can observe that our two techniques significantly outperform the standard baselines in 6 out of 10 possible cases for both topic-relevance and opinion MAP. It is interesting to note that, different from the 2007 dataset, where our approach using the dictionary-based technique did not significantly improve over its base technique, for this dataset, DictProx significantly improved over Dict for 3 baselines in terms of topic-relevance MAP, and for 2 baselines in terms of opinion MAP. More importantly, our approach using OpinionFinder could only significantly improve over its counterpart for baseline5. This observation is in agreement with previous results about the varying effectiveness of these techniques for identifying subjectiveness across different datasets [6].

In order to compare the robustness of our approach with respect to the best performing systems at TREC 2008 and also to better assess its effectiveness across the two datasets considered, Table 3 shows the average improvement of our deployed techniques over all 5 standard baselines in terms of opinion MAP for both 2007 and 2008 topics. The best value in each column is underlined. The median average improvement[1] of the 21 techniques that were applied over all 5 baselines in the TREC 2008 Blog track opinion-finding task is also presented [13]. Their median average improvement on the 2007 topics is included as a reference.

In Table 3, it can be observed that, on average, DictProx does not improve more than Dict on the 2007 topics. OFProx, on the other hand, performs well above the other techniques and is the only one to improve more than the median average improvement on this dataset. Moreover, since the median values on the 2007 topics are those of the 2008 participants and are thus likely to correspond to their training performances on these topics, the average improvement of OFProx on this dataset is even more impressive. A fairer comparison would require the participating systems to run under the same training-testing settings as our approach. On the 2008 topics, DictProx outperforms Dict in terms of average improvement, while OFProx does not improve over OF on average. Furthermore, it can be observed that all techniques improve more than the median TREC average improvement under the fair 2008 training-testing settings.

[1] For each technique, we average its relative improvement across the 5 standard baselines; the final median is then computed across the averages of all 21 techniques.

Overall, these results attest the effectiveness of the OpinionFinder-based techniques when compared to the dictionary-based ones. Nevertheless, it is worth noting that, while the former are much more computationally intensive than the latter [8], integrating the sentence-level proximity feature has a different impact on each technique depending on the underlying query dataset. Indeed, while, on average, OFProx improves over OF on the 2007 dataset, it is only DictProx that is able to improve over its counterpart on the 2008 topics. On average, all techniques provide improvements across all 5 baselines, an achievement only attained by 4 groups in the TREC 2008 Blog track opinion-finding task [13].

6 Conclusions

In this paper, we have proposed a novel approach for blog opinion retrieval aimed at integrating the proximity of query terms to subjective sentences identified from the retrieved documents. We have adapted two techniques known to perform well on the identification of opinionated documents in order to identify subjectiveness at the sentence level.

We have evaluated the effectiveness of our approach using both techniques over 5 standard baselines across two different datasets. Our experiments have shown that our techniques are able to significantly improve over these strongly performing baselines in most cases and perform at least comparably to the top performing systems at the TREC 2007 and 2008 Blog track opinion-finding tasks. Additionally, they can improve over their corresponding base techniques, although this observation was not consistent across the considered datasets. Moreover, we have shown that our approach can perform as well with a high-quality though computationally intensive subjectivity analysis system such as OpinionFinder as with a simple, light-weight approach such as the dictionary-based one.

As a whole, our experiments have demonstrated that our model is feasible and offers a robust performance in the challenging task of improving over a range of statistically different topic-relevance baselines. Furthermore, its generality allows for different techniques for identifying subjective sentences in text to be chosen according to their effectiveness on different datasets and their efficiency.

References

1. Rosenbloom, A.: The blogosphere. Communications of the ACM 47(12), 30–33 (2004)
2. Mishne, G., de Rijke, M.: A study of blog search. In: Lalmas, M., MacFarlane, A., Rüger, S.M., Tombros, A., Tsikrika, T., Yavlinsky, A. (eds.) ECIR 2006. LNCS, vol. 3936, pp. 289–301. Springer, Heidelberg (2006)
3. Macdonald, C., Ounis, I., Soboroff, I.: Overview of the TREC 2007 Blog track. In: Proc. of the 16th Text REtrieval Conference (2007)
4. Ounis, I., de Rijke, M., Macdonald, C., Mishne, G., Soboroff, I.: Overview of the TREC 2006 Blog track. In: Proc. of the 15th Text REtrieval Conference (2006)
5. Ounis, I., Macdonald, C., Soboroff, I.: On the TREC Blog track. In: Proc. of the 2nd International Conference on Weblogs and Social Media, Seattle, WA, USA. AAAI, Menlo Park (2008)

6. Macdonald, C., He, B., Ounis, I., Soboroff, I.: Limits of opinion-finding baseline systems. In: Proc. of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Singapore, pp. 747–748. ACM, New York (2008)
7. Macdonald, C., Ounis, I.: The TREC Blogs 2006 collection: Creating and analysing a blog test collection. Technical Report TR-2006-224, Department of Computing Science, University of Glasgow (2006)
8. He, B., Macdonald, C., He, J., Ounis, I.: An effective statistical approach to blog post opinion retrieval. In: Proc. of the 17th ACM Conference on Information and Knowledge Management, pp. 1063–1072. ACM, New York (2008)
9. Amati, G., Ambrosi, E., Bianchi, M., Gaibisso, C., Gambosi, G.: Automatic construction of an opinion-term vocabulary for ad hoc retrieval. In: Macdonald, C., Ounis, I., Plachouras, V., Ruthven, I., White, R.W. (eds.) ECIR 2008. LNCS, vol. 4956, pp. 89–100. Springer, Heidelberg (2008)
10. Vechtomova, O.: Using subjective adjectives in opinion retrieval from blogs. In: Proc. of the 16th Text REtrieval Conference (2007)
11. Zhou, G., Joshi, H., Bayrak, C.: Topic categorization for relevancy and opinion detection. In: Proc. of the 16th Text REtrieval Conference (2007)
12. He, B., Macdonald, C., Ounis, I.: Ranking opinionated blog posts using OpinionFinder. In: Proc. of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Singapore, pp. 727–728. ACM Press, New York (2008)
13. Ounis, I., Macdonald, C., Soboroff, I.: Overview of the TREC 2008 Blog track. In: Proc. of the 17th Text REtrieval Conference (2008)
14. Wilson, T., Hoffmann, P., Somasundaran, S., Kessler, J., Wiebe, J., Choi, Y., Cardie, C., Riloff, E., Patwardhan, S.: OpinionFinder: A system for subjectivity analysis. In: Proc. of HLT/EMNLP on Interactive Demos (2005)
15. Cacheda, F., Plachouras, V., Ounis, I.: A case study of distributed information retrieval architectures to index one terabyte of text. Information Processing and Management 41(5), 1141–1161 (2005)
16. Metzler, D., Croft, W.B.: A Markov random field model for term dependencies. In: Proc. of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Salvador, Brazil, pp. 472–479. ACM, New York (2005)
17. Srikanth, M., Srihari, R.: Biterm language models for document retrieval. In: Proc. of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Tampere, Finland, pp. 425–426. ACM, New York (2002)
18. Peng, J., Macdonald, C., He, B., Plachouras, V., Ounis, I.: Incorporating term dependency in the DFR framework. In: Proc. of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, The Netherlands, pp. 843–844. ACM, New York (2007)
19. Amati, G.: Probability models for information retrieval based on Divergence From Randomness. PhD thesis, University of Glasgow (2003)
20. Lioma, C., Macdonald, C., Plachouras, V., Peng, J., He, B., Ounis, I.: University of Glasgow at TREC 2006: Experiments in Terabyte and Enterprise tracks with Terrier. In: Proc. of the 15th Text REtrieval Conference (2006)
21. Ounis, I., Amati, G., Plachouras, V., He, B., Macdonald, C., Lioma, C.: Terrier: A high performance and scalable information retrieval platform. In: Proc. of the SIGIR Workshop on Open Source Information Retrieval (2006)