Automatic Construction of an Opinion-Term Vocabulary for Ad Hoc Retrieval

Giambattista Amati[1], Edgardo Ambrosi[2], Marco Bianchi[2],
Carlo Gaibisso[2], and Giorgio Gambosi[3]

[1] Fondazione Ugo Bordoni, Rome, Italy
gba@fub.it
[2] IASI "Antonio Ruberti" - CNR, Rome, Italy
{firstname.lastname}@iasi.cnr.it
[3] University of Tor Vergata, Rome, Italy
gambosi@mat.uniroma2.it

Abstract. We present a method to automatically generate a term-opinion lexicon. We also weight these lexicon terms and use them at real time to boost the ranking with opinionated-content documents. We define very simple models both for opinion-term extraction and document ranking. Both the lexicon model and retrieval model are assessed. To evaluate the quality of the lexicon we compare performance with a well-established manually generated opinion-term dictionary. We evaluate the effectiveness of the term-opinion lexicon using the opinion task evaluation data of the TREC 2007 blog track.

1 Introduction

This work shows how to construct a subjective-word lexicon augmented by term-weights for real-time opinion retrieval. More generally, we address the problem of retrieving documents that contain opinions on a specific topic. Documents, like many posts of web logs (blogs), may contain authors' opinions on a specific subject, and user's information task may consist in retrieving different opinions, like reviews on films, products, books, or more simply people's opinions on public personalities.

The automatic construction of a sentiment and subjective lexicon, and how it can be used for topical opinion are very challenging problems. For example, several machine learning techniques (Naive Bayes, maximum entropy classification, and Support Vector Machines) have been shown to not perform as well on sentiment classification as on traditional topic-based categorization [15]. One of the difficulty of subjective analysis is that sentiment and subjective words distribute quite randomly or more uniformly in the set of relevant documents, while for retrieval or classification models the discriminating words instead occur non-randomly.

To assess the effectiveness of our automatic method we used a dictionary made up of 8221 words built by Riloff, Wiebe and Wilson [16,19]. The words of Riloff et al. dictionary are "clue" words for detecting topical opinions or subjective content, and were collected either manually from different resources or automatically using both annotated and unannotated data. Other opinion term lexicons were created by Mishne [12], and by Esuli and Sebastiani (SentiWordNet) [8]. In particular, Mishne extracted terms

C. Macdonald et al. (Eds.): ECIR 2008, LNCS 4956, pp. 89–100, 2008.

from positive training data using information gain, removing terms appearing also in negative training data, and selecting manually a set of opinion terms.

In this work we show how to generate a sequence of dictionaries

$$Opin\mathbf{V} = Opin\mathbf{V}_1 \supset Opin\mathbf{V}_2 \supset \ldots \supset Opin\mathbf{V}_k \supset \ldots$$

that can be used to topical opinion retrieval. The surprising outcome of this work is that we are able to automatically select and weight the terms of a very small subset $Opin\mathbf{V}_k$ (made up of about 50 words) of the entire dictionary $Opin\mathbf{V}$ (made up of about 12250). This small list of subjective terms performs as good as the entire dictionary in terms of topical opinion retrieval. As a consequence, we are able to perform at real time topical-opinion retrieval with a negligible loss in performance. The reason why we obtain such a high performance is not only due to the technique that singles out the right set of subjective words, but is mainly due to the assigned weights of these subjective words.

This work is developed according to the three following steps:

1. We first define a learning algorithm based on a query expansion technique that selects a set of subjective-term candidates [3]. The selection is based on measuring the divergence of term frequencies in the set of opinionated and relevant documents and in the set of relevant-only documents. High divergence witnesses a potential subjective-term.
2. Then, we assume that the best subjective terms *minimize* the divergence of the within-document term-frequency with respect to the average term-frequency in the set of opinionated and relevant documents. In other words best subjective words spreads over the opinionated and relevant documents more uniformly than the informative words do.
3. We finally introduce a fusion methodology free from any parameter, that combines the content-only ranking with the opinion-only ranking.

Although we use a bag of words approach, we show that topic opinion retrieval performance is very high. Since retrieval, query expansion and ranking merging are obtained by parameter-free functions, our methodology is thus very effective, easy and efficient to implement.

2 Related Work

Topical opinion processing usually is conducted in three steps: extraction of opinion expressions from text (in general seen as a classification problem), document assessment by an opinionated score, and document ranking by opinionated content.

Hatzivassiloglou and McKeown propose data constraints on the semantic orientations of conjoined adjectives, to automatically construct a log-linear regression model predicting whether two conjoined adjectives are of same or different orientation. They further improve the classification of adjectives as positive or negative by defining a graph with orientation links [10]. Agreement on the orientation between adjectives is used as a link, and since positive adjectives tend to be used more frequently than negative ones, one of the two classes that has higher average frequency is classified as having positive semantic orientation.

Using Hatzivassiloglou and McKeown's semantic orientation of adjectives, Turney presents a simple semantic orientation method for phrases based on Mutual Information [6,9] of phrases with adjectives and verbs [18]. A document is classified as recommended if the average semantic orientation of its phrases is positive.

Classification on a whole collection is usually computationally expensive (e.g. Hatzivassiloglou and McKeown's method is NP-complete). A way to reduce the computational cost is to extract information by a topic-driven methodology similar to the query expansion process. For example, Skomowroski and Vechtomova [17] exploit the first-pass retrieval to extract a sample of topic relevant documents, from which co-occurence statistics about adjectives are more efficiently extracted. Nouns are counted when they are in the scope of an adjective, that is adjectives act like modalities or concepts. All subjective adjectives are ranked according to the standard normal score (Z-score) between expected and observed co-occurrences of adjectives with query-terms in the top R retrieved documents. Then, document ranking aggregates different scores, one of them being the opinionated probability of the query terms.

Skomowroski and Vechtomova's work has similar approach to the Local Context Analysis of Xu and Croft [20] who expand the original query taking into account text passages that contain both query-terms (concepts) and expanded-terms. Also Zhang and Yu [21] expand the original query with a list of concepts and a list of expanded words. A classifier for sentences based on Support Vector Machines is trained with some external resources, and then is applied to the set of returned documents. The final document ranking is obtained by removing the documents that do not contain opinions.

There is another approach based on language model that starts with a collection of ternary queries (sentiment, topical, polarity) and collects the statistics in the collection at the sentence level. Their estimate relies on a collection of paired observations, which represent statements for which they know which words are topic and sentiment words. To predict the sentiment value of a new sentence the two word frequencies (in sentence and in collection) are combined by cross-entropy [7].

A central problem for topical opinion document ranking is how to combine ad hoc retrieval scores with additional information on training data, in order to boost the ranks by opinion scores. A simple way to merge scores from different sources of evidence is the use of standard normal scores that has been shown to be very effective in some information tasks [4,17]. Our approach is parameter-free: first we obtain the document ranking by content, then we re-rank the documents taking into account the opinion score and the content rank as combining function.

3 Statistical Analysis of Subjective Terms

The logical representation of languages include three principal constituents: constants c, concepts C and relations R, that roughly, correspond to nouns, adjectives and verbs respectively. A context can be logically represented by $R(C_1(c_1), \ldots, C_k(c_k))$, that is a context is represented by relations among concepts expressed on constants.

However, Information Retrieval has a flat view of objects: the essence of words is their appearance and substance is quantified by probability of occurrence or by means of information theoretic notions like that of information content. It is a matter of fact that nouns provide the highest information content, while adjectives and verbs provide additional information to the context, but bringing less information content.

Our primary goal is to verify some hypotheses on subjective but non-informative terms only by means of information theoretic analysis of term-types and without a direct exploitation of term association and co-occurrence. This simplification will guarantee a faster implementation of opinion analysis. We process terms as for query expansion: we pool all relevant and opinionated documents with respect to all 50 topics of the blog track of TREC 2006, and use the set R of all relevant documents as population and the subset $O \subset R$ of opinionated documents as biased sample. Each term will have four frequencies:

- a relative frequency $\mathbf{p}_c$ in the set $\mathbf{D}$ of all documents;
- a relative frequency $\mathbf{p}_r$ in the set R of relevant documents;
- a relative frequency $\mathbf{p}_o$ in the set O of relevant and opinionated documents;
- a relative frequency $\mathbf{p}_\mathbf{d}$ in the document $\mathbf{d}$.

A dictionary containing weighted terms is automatically generated on the basis of the following considerations:

- Since nouns describe better the content of documents, they possess the highest information content:
$$\text{Inf}(\mathbf{t}) = -\log_2 \text{Prob}(\mathbf{p}_\mathbf{d}|\mathbf{p}_c)$$
The inverse of probability is used to provide the information content of a term in a document $\mathbf{d}$. The main property of Inf is that if $\mathbf{p}_\mathbf{d} \sim \mathbf{p}_c$ then the document is a sample of the collection for the term $\mathbf{t}$ and thus it does not bring information, i.e. $\text{Inf}(t) \sim 0$. $\text{Inf}(\mathbf{t})$ will be used to provide the content score of a term in a document.
- Opinionated terms do not carry information content ($\text{Inf}(\mathbf{t})$ is low). However, they tend to appear more frequently in the opinionated set, O, rather than in the relevant one, R. Therefore, we maximize the opinionated entropy function, $OE(\mathbf{t})$:
$$OE(\mathbf{t}) = -\log_2 \text{Prob}(\mathbf{p}_o|\mathbf{p}_r)$$
to extract possible opinionated terms. On the other hand, information content terms tend to have a similar frequency in both relevant set R and opinionated set O, that is the function $OE(\mathbf{t})$ is minimized for information content terms.
- When nouns are in the scope of adjectives[1], adjectives possibly specify the polarity of opinions. Since verbs link nouns, verbs possibly testify presence of opinions. Concepts[2], adjectives, verbs and adverbs distribute more randomly in the set of opinionated documents. In other words, a high value $OE(\mathbf{t})$ can be due to peaks of frequencies in a restricted number of opinionated documents. The function $OE(\mathbf{t})$

[1] Skomowroski and Vechtomova [17] report that in English a noun follows an adjective the 57% of cases.

[2] According to Heiddeger (1957; Identity and Difference) things are either practical objects or abstracted from their context and reified as ”objects” of our knowledge representation (concepts). Essence of objects, that is the permanent property of things, is the “substance” (understanding), that is the meaning. Nouns mainly represent such objects in our language.

is not robust since it does not consider if the maximization of $OE(\mathbf{t})$ is obtained with a more uniform distribution or not. To filter out noisy terms we use a second information theoretic function (average opinionated entropy, AOE(**t**)) which is the average divergence of document frequency from the expected frequency $\mathbf{p}_o$ in the set of opinionated documents:

$$AOE(\mathbf{t}) = -\frac{1}{|O|}\sum_{\mathbf{d}\in O} \log_2 \text{Prob}(\mathbf{p_d}|\mathbf{p}_o)$$

We will use a very simple approximation of $AOE(\mathbf{t})$ that has not additional cost with respect to the computation of $OE(\mathbf{t})$. The approximation will act as a boolean condition for selecting terms with highest opinion entropy scores $OE(\mathbf{t})$.

The automatically generated dictionary will be further used at retrieval time to re-rank the set of retrieved documents by opinionated scores.

3.1 Distribution of Opinionated Terms in the Set of Opinionated Documents with Respect to Relevant Documents

We have assumed that those terms that occur more often in the set of opinionated documents rather than in the set of relevant documents are possible candidates to bring opinions. To give plausible scores to opinion-bearing terms, we compute an approximation of the opinion entropy $OE(\mathbf{t})$ by means of the asymmetric Kullback-Leibler (KL) divergence computed for all terms in the opinionated set O with respect to the set R of relevant documents, that is

$$OE(\mathbf{t}) = -\log_2 \text{Prob}(\mathbf{p}_o|\mathbf{p}_r) \sim KL(\mathbf{p}_o||\mathbf{p}_r)$$

being $\mathbf{p}_o > \mathbf{p}_r$. We might have used the binomial distribution, or the geometric distribution instead of KL[3] to compute $\text{Prob}(\mathbf{p}_o|\mathbf{p}_r)$, but for the sake of simplicity we prefer to support our arguments with the more intuitive KL measure.

We also anticipated that noise may be caused by some informative terms that appear more densely in a few set of opinionated documents, but the observation of a skewed frequency is mainly due to a more frequent occurrence in the set of documents that are relevant to a given topic. The asymmetric KL divergence is therefore a reliable measure when term-frequency is more randomly or uniformly distributed across all opinionated documents. The noise reduction problem is studied in the following section.

3.2 Distribution of Opinionated Terms in the Set of Opinionated Documents

We want to reduce the noise in opinion-term-selection, that is we want now to filter out those terms that show a distribution of their frequency that is skewed in a few number

[3] KL is an approximation of $\frac{-\log_2 \text{Prob}(\mathbf{p}_o|\mathbf{p}_r)}{\mathbf{TotalFreq(O)}}$ where Prob is the binomial distribution. When weighting terms, the size $\frac{1}{\mathbf{TotalFreq(O)}}$ is a factor common to all words so we may assume that $-\log_2 \text{Prob}(\mathbf{p}_o|\mathbf{p}_r) \sim KL(\mathbf{p}_o||\mathbf{p}_r)$ up to a proportional factor and a small error.

of opinionated documents. A skewed distribution is due to the type of our training data. The opinionated documents are also relevant with respect to a small set of topics (50 queries), and thus it may happen that informative terms might appear more frequently in opinionated documents because a topic may have all relevant documents that are also opinionated, that is when $O(\mathbf{q}) \sim R(\mathbf{q})$: such a situation is not an exception in the blogosphere. In such a case the $OE(\mathbf{t})$ of some non-opinionated terms may be large when compared with the set of all opinionated documents pooled from the set of all topics. We now show how to make a first noise reduction for such cases.

Let $\mathbf{p}_o = \frac{\mathbf{TF}_O}{\mathbf{TotalFreq(O)}}$ be the relative frequency of a term $\mathbf{t}$ in the set of opinionated documents, and $\mathbf{p_d} = \frac{\mathbf{tf}}{\mathbf{l(d)}}$ the relative frequency of the term in the document $\mathbf{d}$. Since the set of opinionated documents O is a large sample of the collection we may set $\mathbf{TotalFreq(O)} = |O| \cdot \bar{\mathbf{l}}$, where $\bar{\mathbf{l}}$ is the average document length and $|O|$ is the number of opinionated documents. The asymmetric KL divergence of the frequency of the term in the opinionated set of document with respect to the prior probability $\mathbf{p}_o = \frac{\mathbf{TF}_O}{|O| \cdot \bar{\mathbf{l}}}$ is:

$$AOE(\mathbf{t}) = \frac{1}{|O|} \sum_{\mathbf{d} \in O} KL(\mathbf{p_d} || \mathbf{p}_o) = \frac{1}{|O|} \sum_{\mathbf{d} \in O} \mathbf{p_d} \cdot \log \frac{\mathbf{p_d}}{\mathbf{p}_o}$$

We have assumed that opinionated terms do not carry information content, and this assumption translates into the assumption that opinion-bearing terms distribute more uniformly in the set of opinionated documents, that is when $\mathbf{p_d} \sim \mathbf{p}_o$, or more generally, when the KL divergence is minimized. If the term distributes uniformly $\mathbf{p_d}$ can be approximated by $\frac{\mathbf{TF}_O}{\mathbf{n_t} \cdot \bar{\mathbf{l}}}$, and we need to *minimize* the function:

$$AOE(\mathbf{t}) \propto - \sum_{\mathbf{d} \in O} \frac{\mathbf{TF}_O}{\mathbf{n_t}} \log \mathbf{n_t} = -\mathbf{n_t} \cdot \frac{\mathbf{TF}_O}{\mathbf{n_t}} \log \mathbf{n_t} = -\mathbf{TF}_O \cdot \log \mathbf{n_t}$$

Since we have to minimize $AOE(\mathbf{t})$ and the approximating expression is negative, and since we may suppose that all terms have a frequency $\mathbf{TF}_O$ of a similar order of magnitude in the set of opinionated documents, we may instead *maximize* the function

$$\log_2 \mathbf{n_t} \propto \mathbf{n_t}$$

where $\mathbf{n_t}$ is the set of opinionated documents containing the term $\mathbf{t}$. We define a term of level k if it appears in at least k relevant and opinionated documents[2]. Therefore the higher the number of documents containing a term, the higher is the probability that the term is opinionated. The larger k is chosen, the less the number of terms that are selected. Therefore, we need to find an optimal level k that generates a vocabulary as small as possible to reduce the computational cost, and in the meantime to be as effective as possible in terms of retrieval performance. The efficiency/effectiveness problems of the automatic generation of an opinionated vocabulary is studied in the following sections.

3.3 Opinion-Term Vocabulary

In summary the information theoretic methodology consists of three steps:

1. Terms with the highest divergence $OE(\mathbf{t})$ between the frequency in the set of opinionated-relevant documents and the frequency in the set of all relevant-only

Table 1. The number of words of the dictionary SCD after the application of the weak Porter stemming is 6,352. The precision and the recall of the automatically generated dictionary $Opin\mathbf{V}_k$ are measured with respect to a semi-manually generated dictionary SCD.

Level	$Opin\mathbf{V}_k \cap SCD$	$Opin\mathbf{V}_k$	Prec.	Rec.	F-Measure
1	2,325	12,263	0.1896	0.3660	0.2498
100	1,528	4,022	0.3800	0.2406	0.2946
250	994	2,504	0.3970	0.1565	0.2245
500	642	1,625	0.3951	0.1011	0.1610
750	466	1,209	0.3854	0.0734	0.1233
1,000	349	927	0.3765	0.0734	0.1228
3,000	77	219	0.3516	0.0121	0.0234
4,000	47	128	0.3672	0.0074	0.0145
6,000	16	42	0.3809	0.0025	0.0050
8,000	5	12	0.4167	0.0008	0.0016

documents are selected, and then weighted by the same opinion-entropy score $OE(\mathbf{t})$. This step generates a list $Cand\mathbf{V}$ of weighted opinion-term candidates.

2. Terms of $Cand\mathbf{V}$ are then filtered. All terms of $Cand\mathbf{V}$ with the lowest average divergence $AOE(\mathbf{t})$ (average divergence between term-frequency in O and the term-frequency within each opinionated-relevant document $\mathbf{d} \in O$), are selected from the list of all terms with positive $OE(\mathbf{t})$ scores. We simply use the minimal number k of opinionated-relevant documents containing the term as fast and effective implementation of the $AOE(\mathbf{t})$ scores.
3. A sequence of weighted dictionaries $Opin\mathbf{V}$ is obtained at different level of k. The optimal level is obtained when the performance is maintained stable while the dictionary size is kept small enough to be used at real-time retrieval.

The $Opin\mathbf{V}_k$ vocabulary is submitted to the system as a standard query and each document obtains an opinionated score. At this aim, in our experiments, we assess the precision of the obtained lexicon and its performance in opinion task retrieval, using a parameter free model of IR (DPH is a variant of the model by Amati [2]) for first pass retrieval and a parameter-free model for query expansion [3]. Using this parameter-free setting for the experiments, we can only concentrate on the methodology to assess the potentiality of the proposed approach. However, other models can be used to enhance initial ranking, because better initial rankings generates better topical opinion rankings.

4 A Computationally Lightweight Algorithm for Topical Opinion Retrieval

The opinionated and relevant document ranking is obtained in three steps:

1. We use the parameter free model DPH as retrieval function to provide the content score of the documents $content\_score(\mathbf{d}||\mathbf{q}) = score_{DPH}(\mathbf{d}||\mathbf{q})$. We obtain a *content rank* for all documents: $content\_rank(\mathbf{d}||\mathbf{q})$.

Table 2. The list of terms of $Opin\mathbf{V}_{6000}$. The table also presents terms of $Opin\mathbf{V}_{6000} \cap SCD$ (italicized terms), terms of $Opin\mathbf{V}_{8000}$ (underlined terms) and $Opin\mathbf{V}_{8000} \cap SCD$ (italicized and underlined terms). A weak Porter stemmer is applied to terms.

am	0,0893	*just*	0,0672	people	0,1094	*view*	0,0139
archive	0,0368	*know*	0,0514	pm	0,1185	wai	0,0303
back	0,0113	last	0,0161	post	0,0326	*want*	0,0395
call	0,0253	left	0,0104	read	0,0293	*well*	0,0187
can	0,0353	*like*	0,0782	*right*	0,0530	who	0,1261
come	0,0193	link	0,0341	sai	0,1124	*will*	0,0070
comment	0,0056	*look*	0,0157	see	0,0350	work	0,0031
dai	0,0247	mai	0,0023	*show*	0,0229	world	0,0286
don	0,0640	*mean*	0,0110	state	0,0049		
first	0,0057	*need*	0,0101	*think*	0,0748		
help	0,0013	now	0,0289	time	0,0407		

2. We submit the entire dictionary $Opin\mathbf{V}$ as a query and weight the set of retrieved documents: $opinion\_score(\mathbf{d}||Opin\mathbf{V}) = score_{DPH}(\mathbf{d}||Opin\mathbf{V})$. The opinionated score with respect to a topic $\mathbf{q}$ is defined as follows[4]:

$$opinion\_score(\mathbf{d}||\mathbf{q}) = \frac{opinion\_score(\mathbf{d}||Opin\mathbf{V})}{content\_rank(\mathbf{d}||\mathbf{q})}$$

   We thus obtain an *opinion rank* for all documents: $opinion\_rank(\mathbf{d}||\mathbf{q})$.
3. We further boost document ranking with the dual function of $opinion\_score(\mathbf{d}||\mathbf{q})$:

$$content\_score^{+}(\mathbf{d}||\mathbf{q}) = \frac{content\_score(\mathbf{d}||\mathbf{q})}{opinion\_rank(\mathbf{d}||\mathbf{q})}$$

   The final topical opinion ranking is obtained re-ranking the documents by $content\_score^{+}(\mathbf{d}||\mathbf{q})$.

5 Experiments and Results

Our experimentation is based on TREC BLOG track dataset [14]. The blog collection was crawled over a period of 11 weeks (December 2005 - February 2006). The total size of the collection amounts to 148 GB with three main different components: feeds (38.6 GB), permalinks (88.8GB), and homepages (20.8 GB). The collection contains

[4] Ranking is a mixture of a normal distribution for relevant documents and an exponential distribution for non-relevant documents [11]. Since the non relevant documents are almost all the documents of the collection for a given query, ranking roughly follows the power law, that is the probability of relevance of a document is inversely proportional to its document rank. Therefore:

$$opinion\_score(\mathbf{d}||\mathbf{q}) \propto opinion\_score(\mathbf{d}||Opin\mathbf{V}) \cdot p(\mathbf{d}||\mathbf{q})$$

spam as well as possibly non-blogs and non-English pages. For our experimentation we considered only the permalinks component, consisting of 3.2 millions of Web pages, each one containing a post and the related comments.

We preprocess data with the aim to remove not English documents from the collection. This goal is achieved by a text classifier, implemented using Lingpipe [1], a suite of Java libraries for the linguistic analysis of human language, and trained using the Leipzig Corpora Collection [5].

We obtain our baseline performing a topic relevance retrieval. For the indexing and retrieval tasks we adopt Terrier [13]. As already stated in Section 3.3, in our experimentation we use DPH, a parameter free retrieval model. This choice has two main consequences: at first we can ignore the tuning issue and focus our efforts on the proposed methodology, evaluating the gain obtained with respect to the baseline; on other hand, all results presented in this Section could be boosted adopting (and properly tuning) a parameter dependent retrieval model.

We use the semi-manual subjectivity clues dictionary [16,19], that we denote here by SCD, to study and enhance the performance of the automatically generated dictionaries, $Opin\mathbf{V}$ in what follows.

Results are shown in Table 3. The first outcome of our work is very surprising: using a set of only 5 subjective and weighted words, that are filtered at the level with $k = 8,000$, we improve both the MAP with respect to relevance (all relevant documents), from 0.3480 of the baseline to 0.3833 (+10%), and the opinionated relevance MAP (only opinionated and relevant documents) from 0.2740 to 0.3128 (+ 14%). Similarly, relevance precision at 10 retrieved documents improves from 0.6480 to 0.7140, while opinionated relevance improves from 0.4480 (0.3031 is the median run of blog track) to 0.5180. It is quite a surprising that a small number of query independent words can improve so largely the quality of ad hoc retrieval. Thus, we may boost both relevance and topical opinion retrieval at real-time with a negligible computational cost.

The best performance values of relevance MAP (0.3938) is obtained with 16 unweighted subjective words (+18% over the median run of TREC 2007 blog track), relevance Precision at 10 (0.7240, +12%) with 349 weighted words, opinionated relevance MAP (0.3213, +33%) with 77 unweighted subjective words, opinionated relevance Precision at 10 (0.5420 , +81%) with 1,528 weighted words. The whole semi manual dictionary SCD containing more than 6,000 of subjective words does not perform as good as its smaller subset $SCD \cap Opin\mathbf{V}_k$ for any level k. This support the idea that it is not the exhaustivity of the dictionary but the subjectivity strength of the words that improves both relevance and topical opinion. More specifically, modalities, conditional sentences or verbs that express possibilities (as the words *can*, *may*) or that relates directly the content to its author (as the words *(I) am*, *like*, *think*, *want*, *agree*) are better predictors of opinions than subjective adjectives. Modal words tend to appear very often in the blogosphere and they alone are almost sufficient to achieve best performance in topical opinion retrieval.

It is worth to note that the $Opin\mathbf{V}_k$ dictionary still contains noisy words due to the fact that we have not used linguistic or lexical tools. As a consequence we did not remove geographical adjectives (e.g. “American”) and other words produced by spam or by blog dependent text in the permalinks (e.g. “post” or “comment”). On the

Table 3. Performance of relevance and topical opinion retrieval by using the semi-manual dictionary SCD and the fully automatic $Opin\mathbf{V}_k$. Test data are from the set 50 queries of the new blog track of TREC 2007. Training data are from the blog track 2006.

	Relevance		Opinion	
Level k	MAP	P@10	MAP	P@10
Baseline	0.3480	0.6480	0.2704	0.4440
Median run of the Blog track 2007	0.3340	0.6480	0.2416	0.3031
SCD	0.3789	0.7000	0.3046	0.5280

$Opin\mathbf{V}_k \cap SCD$, $Opin\mathbf{V}_k$ weighted				
	Relevance		Opinion	
Level k	MAP	P@10	MAP	P@10
1	0.3862	0.7160	0.3173	0.5420
100	0.3862	0.7160	0.3172	**0.5420**
250	0.3864	0.7160	0.3171	0.5380
500	0.3867	0.7160	0.3172	0.5380
750	0.3865	0.7160	0.3168	0.5320
1000	0.3871	**0.7240**	0.3167	0.5380
3000	0.3910	0.7140	**0.3213**	0.5320
4000	0.3909	0.7180	0.3193	0.5300
6000	0.3911	0.7140	0.3204	0.5160
8000	0.3833	0.7140	0.3128	0.5180

$Opin\mathbf{V}_k \cap SCD$, $Opin\mathbf{V}_k$ not weighted				
	Relevance		Opinion	
Level k	MAP	P@10	MAP	P@10
1	0.3801	0.7040	0.3113	0.5340
100	0.3807	0.7100	0.3118	0.5380
250	0.3817	0.7100	0.3126	0.5380
500	0.3825	0.7100	0.3125	0.5340
750	0.3821	0.7000	0.3110	0.5340
1000	0.3836	0.7040	0.3107	0.5340
3000	0.3889	0.7120	0.3135	0.5280
4000	0.3913	0.7120	0.3144	0.5180
6000	**0.3938**	0.7200	0.3123	0.5160
8000	0.3874	0.7120	0.3060	0.4960

Full weighted $Opin\mathbf{V}_k$				
	Relevance		Opinion	
Level k	MAP	P@10	MAP	P@10
1	0.3846	0.7000	0.3080	0.5260
100	0.3848	0.7000	0.3082	0.5260
250	0.3851	0.7000	0.3084	0.5260
500	0.3853	0.7020	0.3083	0.5260
750	0.3856	0.6980	0.3086	0.5220
1000	0.3862	0.7020	0.3103	0.5220
3000	0.3885	0.7040	0.3109	0.5120
4000	0.3879	0.7060	0.3090	0.5080
6000	0.3869	0.7120	0.3103	0.5100
8000	0.3863	0.7060	0.3087	0.5140

$Opin\mathbf{V}_k \cup SCD$, $Opin\mathbf{V}_k$ weighted				
	Relevance		Opinion	
Level k	MAP	P@10	MAP	P@10
1	0.3856	0.7100	0.3168	0.5400
100	0.3856	0.7100	0.3168	0.5400
250	0.3856	0.7100	0.3168	0.5400
500	0.3857	0.7100	0.3170	0.5380
750	0.3860	0.7080	0.3172	0.5360
1000	0.3857	0.7100	0.3165	0.5360
3000	0.3902	0.7140	0.3202	0.5300
4000	0.3903	0.7180	0.3211	0.5380
6000	0.3899	0.7140	0.3205	0.5360
8000	0.3871	0.7160	0.3166	0.5380

other hand, removing words is a challenging task, since $Opin\mathbf{V}_k$ contains words that are exclamations, slang or vulgar words that express emotions or opinions but that do not belong to a clean dictionary like SCD. Furthermore some words are missing (e.g. "good" or "better") because the collection has been indexed using the default stopword list provided by the Terrier framework.

6 Conclusions

We have automatically generated a dictionary of subjective words and we have introduced a method to weight the words of the dictionary through information theoretic measures for topical opinion retrieval. In contrast to term-association or co-occurrence techniques, we have used the training collection as a bag of words. We have first learned all possible subjective words candidates by measuring the divergence of opinionated term-frequencies from only-relevant term-frequencies. Then, we have made the assumption that the best (most discriminating) subjective words are the most frequent ones, and that they distribute non-randomly in the set of opinionated documents. Following this hypothesis, we built a sequence of refined dictionaries, each of them shows to keep almost unaltered the performance for both retrieval tasks (relevance and opinionated relevance), up to the limit point of using a very small number of words of the dictionary. Our opinionated relevance ranking formula is also very robust and does not need any parameter tuning or learning from relevance data. Because of the small size of these dictionaries, we may boost opinionated and relevant documents at real-time with a negligible computational cost. Further refinements of the dictionary are possible, for example using lexical or other external resources. Also minimization of the average divergence $AOE(\mathtt{t})$, that filters out good subjective words, can be computed more accurately than the first approximation we have used for these experiments.

References

1. Alias-i. Lingpipe named entity tagger, `http://www.alias-i.com/lingpipe/`
2. Amati, G.: Frequentist and Bayesian approach to Information Retrieval. In: Lalmas, M., MacFarlane, A., Rüger, S.M., Tombros, A., Tsikrika, T., Yavlinsky, A. (eds.) ECIR 2006. LNCS, vol. 3936, pp. 13–24. Springer, Heidelberg (2006)
3. Amati, G., Carpineto, C., Romano, G.: Query difficulty, robustness, and selective application of query expansion. In: McDonald, S., Tait, J.I. (eds.) ECIR 2004. LNCS, vol. 2997, pp. 127–137. Springer, Heidelberg (2004)
4. Amati, G., Carpineto, C., Romano, G.: Merging xml indices. In: Fuhr, N., Lalmas, M., Malik, S., Szlávik, Z. (eds.) INEX 2004. LNCS, vol. 3493, pp. 253–260. Springer, Heidelberg (2005)
5. Biemann, C., Heyer, G., Quasthoff, U., Richter, M.: The leipzig corpora collection - monolingual corpora of standard size. In: Proceedings of Corpus Linguistic 2007, Birmingham, UK (2007)
6. Church, K.W., Hanks, P.: Word association norms, mutual information, and lexicography. In: Proceedings of the 27th. Annual Meeting of the Association for Computational Linguistics, pp. 76–83. Association for Computational Linguistics, Vancouver, B.C (1989)
7. Eguchi, K., Lavrenko, V.: Sentiment retrieval using generative models. In: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, July 2006, pp. 345–354. Association for Computational Linguistics, Sydney, Australia (2006)
8. Esuli, A., Sebastiani, F.: SentiWordNet: A publicly available lexical resource for opinion mining. In: Proceedings of LREC-06, the 5th Conference on Language Resources and Evaluation (2006)
9. Fano, R.M.: Transmission of Information: A Statistical Theory of Communications. MIT Press, Cambridge, Wiley, New York (1961)
10. Hatzivassiloglou, V., McKeown, K.: Predicting the semantic orientation of adjectives. In: acl97, pp. 174–181 (1997)
11. Manmatha, R., Rath, T., Feng, F.: Modeling score distributions for combining the outputs of search engines. In: SIGIR 2001: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 267–275. ACM, New York (2001)

12. Mishne, G.: Multiple ranking strategies for opinion retrieval in blogs. In: The Fifteenth Text REtrieval Conference (TREC 2006) Proceedings (2006)
13. Ounis, I., Amati, G., Plachouras, V., He, B., Macdonald, C., Lioma, C.: Terrier: A High Performance and Scalable Information Retrieval Platform. In: Proceedings of ACM SIGIR'06 Workshop on Open Source Information Retrieval (OSIR 2006) (2006)
14. Ounis, I., de Rijke, M., Macdonald, C., Mishne, G., Soboroff, I.: Overview of the trec-2006 blog track. In: Proceedings of the Text REtrieval Conference (TREC 2006), National Institute of Standards and Technology (2006)
15. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up?: sentiment classification using machine learning techniques. In: EMNLP 2002: Proceedings of the ACL-02 conference on Empirical methods in natural language processing, pp. 79–86. Association for Computational Linguistics, Morristown, NJ, USA (2002)
16. Riloff, E., Wiebe, J.: Learning extraction patterns for subjective expressions. In: Proceedings of the 2003 conference on Empirical methods in natural language processing, pp. 105–112. Association for Computational Linguistics, Morristown, NJ, USA (2003)
17. Skomorowski, J., Vechtomova, O.: Ad hoc retrieval of documents with topical opinion. In: Amati, G., Carpineto, C., Romano, G. (eds.) ECIR 2007. LNCS, vol. 4425, pp. 405–417. Springer, Heidelberg (2007)
18. Turney, P.: Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In: acl2002, pp. 417–424 (2002)
19. Wilson, T., Wiebe, J., Hoffmann, P.: Recognizing contextual polarity in phrase-level sentiment analysis. In: Proceedings of HLT-EMNLP (2005)
20. Xu, J., Croft, W.B.: Query expansion using local and global document analysis. In: Proceedings of ACM SIGIR, Zurich, Switzerland, August 1996, pp. 4–11 (1996)
21. Zhang, W., Yu, C.: Uic at trec 2006 blog track. In: The Fifteenth Text REtrieval Conference (TREC 2006) Proceedings (2006)