FUB, IASI-CNR and University of Tor Vergata at TREC 2008 Blog Track

Giambattista Amati[1], Giuseppe Amodeo[1], Marco Bianchi[2], Carlo Gaibisso[2], and Giorgio Gambosi[3]

[1] Fondazione Ugo Bordoni, Rome, Italy gba@fub.it
[2] Istituto di Analisi dei Sistemi ed Informatica "A. Ruberti"-CNR, Rome, Italy carlo.gaibisso@iasi.cnr.it
[3] Mathematics Department of University of Tor Vergata, Rome, Italy gambosi@mat.uniroma2.it

1 Introduction

We take part in the opinion and polarity retrieval tasks of the blog track.

A test collection, called Blog06, was created for the blog track in 2006 [4] with three main different components: feeds, permalinks and home-pages. The collection contains spam as well as possibly no blogs and no english pages. For our experimentation only permalinks have been taken into consideration, consisting of 3.2 million of Web pages for a total of 88.8GB, each one containing a post and its related comments.

The evaluation metrics are precision/recall based [4], the Mean Average Precision (MAP) and R-Precision (RPrec), but we also focused on Precision at 10 (P@10), due to its relevance in evaluating the effectiveness of Web search engines [5] [3].

As in 2007, we based our approch on the costruction of ad-hoc weighted dictionaries, containing terms assumed to be used to express a sentiment. The weight is a measure *of how much* sentiment the term expresses.

To automatically construct our dictionaries, we assumed that "*opinion-bearing*" words distribute more randomly in the set of opinionated documents than semantic-bearing terms, but less randomly than not-informative terms.

As a consequence, we relyed on two theoretic measures. The first of them was based on a Divergence From Randomness (DFR) model and defined the weight of each term within an opinionated document, conseguently identifing the set of terms candidate to appear in the vocabularies. The other one, was based on entropy maximization in the set of all relevant and opinionated documents and defined the final content of the dictionaries and the weights of their terms.

By these dictionaries, we first reranked the set of documents relevant to a topic on the basis of the quantity of opinion they express, and then extract two new rankings according to the polarity of the expressed sentiment.

All these phases are detailed described in Sections 2, 3, 4, 5 and 6. Finally, in Section 7 we report and discuss on the experimentation activity and results. Finally, a brief analysis of our results is present in 8.

2 Data preprocessing

As in 2007 [5], data preprocessing mainly consisted in trying to remove no english documents from the collection through LingPipe [1]. In our intention the tool should also succed in detecting some of the spam. A deeper analysis, than those conducted in 2007, of the effectiveness of this approch, revealed that a consistent fraction of relevant documents were wrongly identified as written in a language other than english.

Unfortunately we hadn't enough time to test alternative training modalities of the LingPipe or to evaluate complete different approches to solve the problem. Thus, we have been forced to deal with the original collection, spam and no english documents included.

3 Topic relevance retrieval

For the retrieval of the documents relevant to a topic, we basically followed the same approch adopted in 2007, with only few exceptions: we did not rely on the distributed implementation of Terrier [2] to build our indexes, while DFRee1, a parameter free retrieval model, has been adopted instead of DPH. The stemmimg modalities and the choice of the parametric PL2 model, with c set to 9, stayed unaffected.

Table 1 shows the values of the MAP, the R-Prec and the P@10 for the topic relevance retrieval baselines we submitted to TREC 2008 (BL_DFRee and BL_PL2c9 respectively for the DFRee1 and PL2 retrieval models). Together with the same values for the baselines provided by NIST (BL1, BL2, BL3, BL4, BL5).

4 Automatic costruction of ad-hoc dictionaries

Our approch is based on the costruction of three ad-hoc weighted dictionaries: one for the opinion retrieval, $Opin\mathbf{D}$, and the other two for the polarity detection, $Pos\mathbf{D}$ and $Neg\mathbf{D}$. Before entering the details of the construction, let us introduce a little bit of notation. Let:

- $\mathcal{C}$ denote the collection of documents;
- $\mathcal{R} \subseteq \mathcal{C}$ denote the set of documents relevant to a topic;
- $\mathcal{O} \subseteq \mathcal{R}$ denote the set of documents relevant to the same topic, expressing an opinion on it;

In automatically constructing $Opin\mathbf{D}$ [3] [5], we assumed that:

- content-bearing words maximize the probability P of observing the posterior probability of occurrence in $\mathcal{O}$, given the prior probability of occurrence in $\mathcal{R}$.
- opinion-bearing words, instead, minimize the same probability. The weight of an opinion-bearing word is provided by a DFR model, and, as a consequence, a word is assumed to express an opinion iff it maximizes the value of the computed divergence.

– best opinion-bearing words also maximize the entropy in $\mathcal{O}$. In our assumption, the approach we adopted to maximize entropy is to select the terms with highest divergence, that, at the same time, belong to a *large enough* number of opinionated documents.

Starting from our assumptions, $\mathcal{R}$ has been identified as the set of documents recognized as relevant by TREC 2006-2007, those labeled 1, 2, 3 or 4 in the provided qrels; while $\mathcal{O}$ as the set of opinionated ones, those labeled 2, 3 or 4 in the same qrels [4].

The DFree1 DFR has been adopted to identify the set of opinion-bearing words. The set of best opinion-bearing words is then obtained as follows: a sequence of candidate dictionaries $D_1 \supseteq D_2 \supseteq \cdots \supseteq D_k$, with D_1 coinciding with the set of opinion-bearing words, has been computed such that $\forall i = 1, \ldots, n\ D_i = \{t \in D_1 \wedge df_t \geq i\}$ where df_t is the document frequency of term t in $\mathcal{O}$ [3] [5].

As result, a generic k level dictionary contains all opinion-bearing terms occurring in at least k documents in $\mathcal{O}$. Our final goal was to find the maximum value of k, say $\overline{k}$, that keeps the retrieval performance *stable*, when compared with those obtained by D_k, with $k \leq \overline{k}$; and maintains the dictionary size small engough to be computationally effective. The value of k best fitting our needs has been tentatively fixed to $1,000$.

$Pos\mathbf{D}$ and $Neg\mathbf{D}$ respectively are analogously determined: all the above assumptions and considerations still hold if $\mathcal{R}$ is substituted by $\mathcal{O}$, and $\mathcal{O}$ by $\mathcal{O}^+$ and $\mathcal{O}^-$, respectively, where $\mathcal{O}^+$ (resp. $\mathcal{O}^-$) denotes the set of documents expressing a positive (resp., negative) opinion.

This time $\mathcal{O}$ has been identified as the set of documents recognized as positively and negatively opinionated by TREC 2006-2007, those labeled 2 or 4 in the provided qrels. The value of k best fitting our needs has been tentatively fixed to 500 for $Pos\mathbf{D}$, and to 100 for $Neg\mathbf{D}$. Since weights assigned to terms appeared to be significantly *dissimilar* between the two dictionaries, the weights of each dictionary have been normalized to the highest value inside of the dictionary itself.

5 Opinionated relevance retrieval

Opinionated and relevant documents was ranked, for each query $\mathbf{q}$, in three steps:

1. a topic retrieval step was accomplished, as decribed in section 3: a new rank, say $content\_rank(\mathbf{d}||\mathbf{q})$, was assigned to each document d, depending on the score, say $content\_score(\mathbf{d}||\mathbf{q})$, assigned to it by the adopted DFR model;
2. a new query, maden by all the terms in $Opin\mathbf{D}$, weighted by their respective weights, was submitted: a new score was obtained for each document d, say $opinion\_score(\mathbf{d}||Opin\mathbf{D})$. A new rank, say $opinion\_rank(\mathbf{d}||\mathbf{q})$, for each document d was then obtained on the basis of $opinion\_score(\mathbf{d}||\mathbf{q}) = opinion\_score(\mathbf{d}||Opin\mathbf{V})/content\_rank(\mathbf{d}||\mathbf{q})$;

3. the final ranking was obtained by furtherly boosting the rank assigned to each document d, say $content\_score^{+}(\mathbf{d}||\mathbf{q})$, as follows: $content\_score^{+}(\mathbf{d}||\mathbf{q}) = content\_score(\mathbf{d}||\mathbf{q})/opinion\_rank(\mathbf{d}||\mathbf{q})$.

6 Polarity Recognition

Polarity recognition is accomplished with an approach similar to that adopted for the opinionated relevance retrieval. The starting point is the opinion ranking determined according to the modalities described in section 5. The computation of the polarity rank is based on the weights assigned to the terms in $Pos\mathbf{D}$ and in $Neg\mathbf{D}$. The final polarity score of a document is obtained by subtracting to its positive polarity score its negative one. If the final score is greater than zero, the document is considered as expressing a positive opinion; a negative opinion, otherwise. Finally if this score is *close* to zero, we consider the document as not sufficently polarized.

7 Tests and results

We first of all generate our baselines, one for each topic of interest, by ranking the documents according to the content they bear: table 1 shows the mean of the values of the topic relevance MAP, R-Prec and P@10 for our baselines, rows BL_DFRee and BL_PL2c9 for the DFRee1 and PL2 retrieval models, together with the same values for the baselines provided by the NIST. As shown by the table, in no case we succeeded to improve the baselines of reference.

Next, each baseline is re-ranked according to the *quantity of* opinion its documents bear. These new rankings will be referred to as *opinion based rankings*. To asses the effectiveness of our approach, we first of all investigate its impact on the baselines: we compared the MAP and the R-Precision values of table 1, with the corrisponding values for the opinion based rankings generated using the DFRee1 and PL2 models, shown by tables 2 and 3, respectively. These tables also show the results of the comparison.

We then compared the MAP and the R-Prec values for the opinion relevance of the baselines, shown by table 4, with the corrisponding values for the opinion based rankings generated using the DFRee1 and PL2 models, shown by tables 5 and 6, respectively. Table 7 shows the results of this comparison.

Furthermore, for each topic we have been given the medians of the opinion relevance MAP and R-Prec for all the runs submitted by all the participants. We compute the means of this values, 0.3050 for the MAP and 0.3651 for the R-Prec, and compare them with our results, as shown by tables 5 and 6.

Finally, documents in each of the opinion based rankings are filtered according to the positive (resp., negative) polarity of the opinion they bear, resulting in two new rankings, one of documents expressing a positive opinion, the other of documents expressing a negative one. It is worth noting that the sets of documents in these rankings, do not intersect and are a subset of the documents appearing in the original opinion based ranking. This implies that the MAP and

R-Prec values for these rankings can not be directly compared with those for the opinion based rankings.

As a consequence we limited ourselves to investigate the effectiveness of our approach by comparing the medians of the polarity relevance MAP, i.e. 0.1151, and R-Prec, i.e. 0.1624, for the runs submitted by all participans with our results, as shown by tables 8 and 9.

8 Conclusion

By our experiments we confirmed that our approach to the opinion retrieval is really effective and robust, also in absence of ad hoc solutions for the detection of spam and of no english documents. The Dfree1 model has proved itself to overperform the PL2 one. As concerns the polarity detection, we failed in achieving acceptable results. We think that the main motivation of this failure has reference to the scarce effectiveness of our sentimental dictionaries in properly classifying documents. May be the adoption of an approach based on passage retrieval could be the proper solution to this problem.

References

1. *Alias-i. Lingpipe named entity tagger. In http://www.alias-i.com/lingpipe/.*
2. I. Ounis, G. Amati, V. Plachouras, B. He, C. Macdonald, and C. Lioma. Terrier: A High Performance and Scalable Information Retrieval Platform. In *Proceedings of ACM SIGIR'06 Workshop on Open Source Information Retrieval (OSIR 2006)*, 2006.
3. G. Amati,E. Ambrosi, M. Bianchi, C. Gaibisso, and G. Gambosi Automatic Construction of an Opinion-Term Vocabulary for Ad Hoc Retrieval In *Proceedings of ECIR 2008, LNCS 4956, pp. 89–100*, 2008.
4. C. Macdonald, I. Ounis, University of Glasgow Overview of the TREC 2007 Blog Track In *Proceedings of The Sixteenth Text REtrieval Conference (TREC 2007)*, 2008.
5. G. Amati,E. Ambrosi, M. Bianchi, C. Gaibisso, and G. Gambosi FUB, IASI-CNR and University of Tor Vergata at TREC 2007 Blog Track In *Proceedings of The Sixteenth Text REtrieval Conference (TREC 2007)*, 2008.

Run	MAP	R-Prec	P@10
BL_DFR	0.3195	0.3756	0.6307
BL_PL2c9	0.3287	0.3838	0.6313
BL1	0.3701	0.4156	0.7307
BL2	0.3382	0.3831	0.7000
BL3	0.4244	0.4573	0.7220
BL4	0.4776	0.5092	0.7867
BL5	0.4424	0.4868	0.7793

Table 1.

Run	MAP	R-Prec	$\Delta_{\%}$ MAP	$\Delta_{\%}$ R-Prec
FIUDFRDFR	0.3483	0.4036	9.0%	7.4%
FIUPL2c9DFR	0.3546	0.4089	7.9%	6.5%
FIUBL1DFR	0.4073	0.4529	10.0%	9.0%
FIUBL2DFR	0.3401	0.3866	0.6%	0.9%
FIUBL3DFR	0.4437	0.4753	4.5%	3.9%
FIUBL4DFR	0.4854	0.5153	1.6%	1.2%
FIUBL5DFR	0.0975	0.0799	-78.8%	-83.6 %

Table 2.

Run	MAP	R-Prec	$\Delta_{\%}$ MAP	$\Delta_{\%}$ R-Prec
FIUDFRPL2c9	0.3451	0.3990	8.0%	6.2 %
FIUPL2PL2c9	0.3495	0.4012	6.3%	4.5%
FIUBL1PL2c9	0.4082	0.4518	10.3%	8.7%
FIUBL2PL2c9	0.3284	0.3735	-2.9%	-2.5%
FIUBL3PL2c9	0.4416	0.4722	4.0%	3.3%
FIUBL4PL2c9	0.4775	0.5070	0.0%	-0.4%
FIUBL5PL2c9	0.0995	0.0858	-77.51%	-82.4%

Table 3.

Run	MAP	R-Prec	P@10
BL_DFR	0.2416	0.3028	0.4547
BL_PL2c9	0.2462	0.3041	0.4647
BL1	0.2639	0.3189	0.4753
BL2	0.2657	0.3189	0.5287
BL3	0.3201	0.3647	0.5387
BL4	0.3543	0.3979	0.5580
BL5	0.3147	0.3709	0.5307

Table 4.

Run	MAP	R-Prec	$\Delta_{\%}$ MAP	$\Delta_{\%}$ R-Prec
FIUDFRDFR	0.2745	0.3379	14.2%	10.5%
FIUPL2c9DFR	0.2770	0.3370	16.3%	12.0%
FIUBL1DFR	0.3033	0.3569	33.5%	24.0%
FIUBL2DFR	0.2774	0.3287	11.5%	5.9%
FIUBL3DFR	0.3436	0.3883	45.5%	30.2%
FIUBL4DFR	0.3760	0.4175	59.2%	41.1%
FIUBL5DFR	0.0607	0.0360	-68.0%	-78.1%

Table 5.

Run	MAP	R-Prec	$\Delta_{\%}$ MAP	$\Delta_{\%}$ R-Prec
FIUDFRPL2c9	0.2735	0.3371	13.1%	9.3%
FIUPL2PL2c9	0.2752	0.3309	14.6%	9.9%
FIUBL1PL2c9	0.3055	0.3577	33.8%	23.7%
FIUBL2PL2c9	0.2688	0.3164	7.7%	2.3%
FIUBL3PL2c9	0.3438	0.3843	44.8 %	29.3%
FIUBL4PL2c9	0.3725	0.4159	56.6%	38.9%
FIUBL5PL2c9	0.0614	0.0393	-67.4%	-76.5%

Table 6.

Run	$\Delta_{\%}$ MAP	$\Delta_{\%}$ R-Prec	$\Delta_{\%}$ P@10
BL_DFR	12.0% — 11.7%	10.4% — 10.2%	12.4% — 12.8%
BL_PL2c9	12.5% — 10.5%	10.8% — 8.1%	10.2% — 9.4%
BL1	13.0 % — 13.6 %	10.6% — 10.8%	11.5% — 13.4%
BL2	4.2% — 1.2%	3.0% — -0.8%	3.3% — 3.4%
BL3	6.8% — 6.9%	6.1% — 5.1%	5.7% — 6.6%
BL4	5.8% — 4.9%	4.7% — 4.3%	5.3% — 6.5%
BL5	-80.7% — -80.5%	-90.3% — -89.4%	-92.3% — -91.9%

Table 7.

Run	MAP	R-PRec	$\Delta_{\%}$ MAP	$\Delta_{\%}$ R-Prec
FIUpDFRDFR	0.0569	0.1058	-40%	-22%
FIUpPL2DFR	0.0561	0.1076	-51%	-34%
FIUpBL1DFR	0.0686	0.1269	-41%	-21%
FIUpBL2DFR	0.0560	0.1074	-31%	-13%
FIUpBL3DFR	0.0681	0.1277	-86%	-82%
FIUpBL4DFR	0.0793	0.1406	-51%	-35%
FIUpBL5DFR	0.0158	0.0285	-51%	-34%

Table 8.

Run	MAP	R-PRec	$\Delta_{\%}$ MAP	$\Delta_{\%}$ R-Prec
FIUpDFRDFR	0.0484	0.0821	-28%	-20%
FIUpPL2DFR	0.0481	0.0802	-24%	-17%
FIUpBL1DFR	0.0481	0.0801	14%	12%
FIUpBL2DFR	0.0507	0.0831	-4%	-2%
FIUpBL3DFR	0.0760	0.1124	-70%	-76%
FIUpBL4DFR	0.0640	0.0981	-27%	-18%
FIUpBL5DFR	0.0198	0.0238	-28%	-20%

Table 9.