Using Opinion-Based Features to Boost Sentence Retrieval

Ronald T. Fernández
Grupo de Sistemas Inteligentes
Departamento de Electrónica y Computación
Universidad de Santiago de Compostela, Spain
ronald.teijeira@usc.es

David E. Losada
Grupo de Sistemas Inteligentes
Departamento de Electrónica y Computación
Universidad de Santiago de Compostela, Spain
david.losada@usc.es

ABSTRACT

Opinion mining has become recently a major research topic. A wide range of techniques have been proposed to enable opinion-oriented information seeking systems. However, little is known about the ability of opinion-related information to improve regular retrieval tasks. Our hypothesis is that standard retrieval methods might benefit from the inclusion of opinion-based features. A sentence retrieval scenario is a natural choice to evaluate this claim. We propose here a formal method to incorporate some opinion-based features of the sentences as query-independent evidence. We show that this incorporation leads to retrieval methods whose performance is significantly better than the performance of state of the art sentence retrieval models.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H3.3 Information Search and Retrieval.

General Terms

Theory, Experimentation

Keywords

Opinion mining, Sentence retrieval.

1. INTRODUCTION

Opinion mining deals with the computational treatment of opinions, sentiment and subjectivity in texts [7]. A classification module, which consists of detecting opinions and estimating their polarity, is usually a core component of the sentiment-aware systems. Document passages (e.g. sentences) may be classified following their opinionated nature. Additionally, subjective material can be classified as expressing either an overall positive or an overall negative opinion (polarity classification). However, the role of opinionated information in standard retrieval tasks is currently unknown. We argue here that opinion-based features are a valuable component that can enhance state of the art retrieval algorithms. Our intuition is that users might be particularly interested in subjective material and, therefore, promoting subjective pieces of information might become a way to improve retrieval performance. We test this assumption in the context of sentence retrieval (SR). This is convenient because sentences are compact pieces of text that cover a narrow topic and can be reasonably classified as subjective or objective.

In this paper, a formal study will be conducted to check whether opinion-based features can be used to improve SR performance. We will follow a formal methodology [2] to incorporate this query-independent evidence into existing retrieval models. This helps to study properly the combination of a query-sentence matching score with a query-independent score computed from opinion-based features.

The rest of the paper is organized as follows. Section 2 presents the opinion-based features and the software utilized to estimate them. The methodology followed to combine SR scores with opinion-based scores is explained in Section 3. Section 4 reports the experiments and analyzes their outcomes. The paper ends with Section 5, where we expose the conclusions of our study.

CIKM'09, November 2–6, 2009, Hong Kong, China.

2. OPINION-BASED FEATURES

To extract the opinion-based features associated to every sentence we use a highly effective opinion mining software, which we describe below.

OpinionFinder [9] is a state of the art subjectivity detection system [7, 8] that works as follows. First, the text is processed using part-of-speech tagging, name entity recognition, tokenization, stemming, and sentence splitting. Next, a parsing module builds dependency parse trees where subjective expressions are identified using a dictionary-based method. This is powered by Naive Bayes classifiers that are trained on sentences automatically generated from unannotated data.

Sentences are classified as subjective or objective (or unknown if it cannot determine the nature of the sentence). Two classifiers are implemented by OpinionFinder: accuracy classifier and precision classifier. The first one yields the highest overall accuracy. It tags each sentence as either subjective or objective. The second classifier optimizes precision at the expense of recall. It classifies a sentence as subjective or objective only if it can do so with confidence. Furthermore, OpinionFinder marks various aspects of the subjectivity in the sentences, including the words that are estimated to express positive or negative sentiments.

We work in this paper with the following opinion-based features: F_{subj}, the subjective nature of the sentence (this

is a binary value, which is 1 when the sentence is classified as subjective, and 0 otherwise); F_{pos}, the number of positive terms in a sentence; F_{neg}, the number of negative terms in a sentence; and F_{opt} the number of opinionated terms in a sentence. Observe that all these features are discrete and the last three features range from zero to the total number of sentence terms.

In the following section, we present formal methods that define query-independent relevance weights using these opinion-based features.

3. COMBINING CONTENT MATCH AND OPINION-BASED SCORES

Sentence Retrieval (SR) consists of finding sentences that are relevant to a given information need. This is usually done by retrieving, first, a set of potentially relevant documents and, next, running a SR algorithm against the sentences in these documents. This task is required in a wide range of Information Retrieval applications, such as summarization, question-answering, etc.

Traditional SR methods proposed in the literature are often based on a regular matching between the query and every sentence. A vector-space approach, the tfisf model [1], is a simple but very effective SR method [1, 6]. It is parameter-free but performs at least as well as tuned SR methods based on Language Models or BM25 [5]. Along this work, we use tfisf as our SR baseline.

As argued above, we hypothesize here that SR methods can be further improved by guiding the retrieval process towards opinionated sentences. A natural solution is to define query-independent weights that modify the SR score provided by tfisf. To do this adjustment we apply FLOE, a formal methodology designed by Craswell et al [2]. FLOE (feature's log odds estimate) is a density analysis method for modelling the shape of the transformation required to incorporate a query-independent feature into an existing retrieval model. It is a powerful method that finds good functions to transform feature values into relevance scores, without assuming independence between the feature and the baseline.

Next, we describe how to apply FLOE to find a proper score adjustment in our SR scenario.

3.1 Score Adjustment

Our aim is to rank sentences (S) according to the probability that they are relevant (R), $p(R|S)$. We can rewrite this in a log-odds way which preserves the rank order with respect to the query (Q) [2]:

$$p(R|S) \overset{Q}{\propto} \log \frac{p(R|S)}{p(\bar{R}|S)} \overset{Q}{\propto} \log \frac{p(S|R)}{p(S|\bar{R})} \quad (1)$$

Sentences can be considered to have two components: a content match component (M) and a query-independent (or static score) component (I). Given these two components, equation 1 can be rewritten as:

$$\log \frac{p(S|R)}{p(S|\bar{R})} = \log \frac{p(M,I|R)}{p(M,I|\bar{R})} = \log \frac{p(M|R)}{p(M|\bar{R})} + \log \frac{p(I|M,R)}{p(I|M,\bar{R})} \quad (2)$$

A regular matching function, such as tfisf, can play the role of the first addend [2]:

$$\log \frac{p(S|R)}{p(S|\bar{R})} \overset{Q}{\propto} tfisf + \log \frac{p(I|M,R)}{p(I|M,\bar{R})} \quad (3)$$

Assuming that components I and M are independent, the query-independent adjustment is: $\log \frac{p(I|R)}{p(I|\bar{R})}$.

Since the number of relevant sentences is very small compared to the number of sentences in the collection, the correct adjustment under this independence assumption can be approximated as:

$$indep(I,R) = \log \frac{p(I|R)}{p(I)} \quad (4)$$

Figure 1 shows the curves for the probabilities $p(I|R)$ and $p(I)$ in one of our training collections (TREC 2003 Novelty Track dataset)[1]. The F_{neg}, F_{pos} and F_{opt} graphs were smoothed by applying a shape preserving interpolation[2].

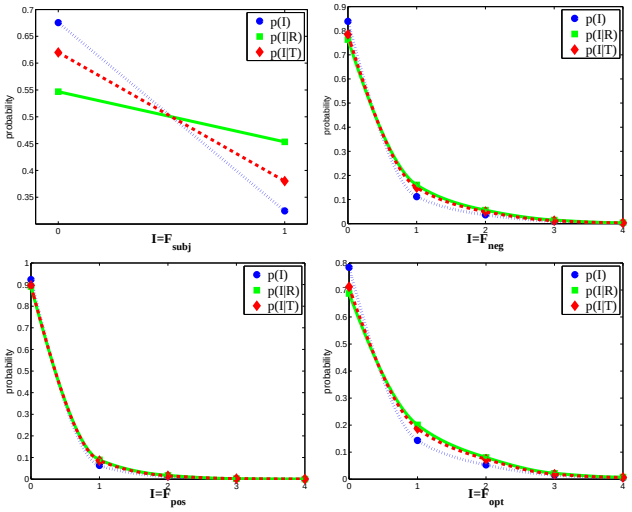

Figure 1: $p(I)$, $p(I|R)$ and $p(I|T)$ for the opinion-based features.

With the binary feature F_{subj}, $p(F_{subj} = 1|R) > p(F_{subj} = 1)$ and $p(F_{subj} = 0|R) < p(F_{subj} = 0)$, i.e. the percentage of subjective sentences in the relevant set is higher than the overall percentage of subjective sentences in the collection. This supports our belief that promoting subjective sentences might be a way to achieve better performance because the distribution of subjective sentences is larger in the set of relevant sentences compared to the entire collection. This would suggest that using F_{subj} as an informative prior would improve performance by providing an initial weight towards subjectivity before the query-dependent information is considered in the ranking model.

The tendencies with F_{neg}, F_{pos} and F_{opt} are less obvious but, still, we can identify some trends. $p(I|R)$ tends to be greater than $p(I)$ when $I \geq 1$ (this is particularly apparent with $I = 1$). In contrast, $p(I)$ tends to be greater than $p(I|R)$ when $I = 0$. Again, this evidence seems to indicate that these polarity terms tend to appear more in the relevance sets than in the collection. Note also that there is little

[1]Similar plots were obtained for all training collections described in Section 4.

[2]Craswell et al [2] applied kernel density smoothing but we do not need it here because our features are discrete. The smoothed curves are simply built to have a clearer view of how probability evolves with the feature's value. Nevertheless, the features take integer values and, therefore, the graphs should only be analyzed for the points whose x coordinate is an integer.

distinction between $p(I)$ and $p(I|R)$ with F_{pos}. We therefore anticipate that positive terms will be a less valuable indicator of relevance.

The indep score represents the adjustment suggested under the independence assumption (baseline and opinion-based features are independent). The indep curves, which are shown in Figure 2, suggest that we need to give more weight to sentences with opinionated material ($F_{subj} = 1$ or F_{pos}, F_{neg}, $F_{opt} \geq 1$). This adjustment even suggests to remove some weight from sentences with no opinionated information ($F_{subj} = 0$ leads to a negative weight).

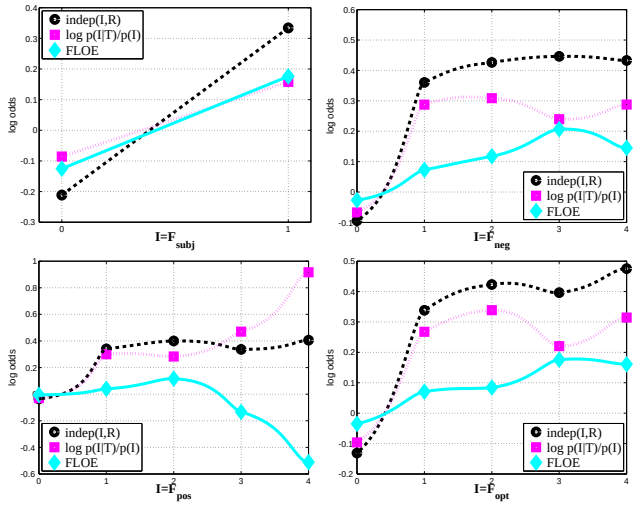

Figure 2: FLOE adjustment for the four features. Adding FLOE to log $\frac{p(I|T)}{p(I)}$ we get the ideal adjustment, indep.

This approach would perform well if the baseline and the query-independent scores would be actually independent. However, it might be the case that the baseline already retrieves enough subjective material. To account for dependency, Craswell et al [2] defined FLOE, which we describe next.

3.2 FLOE: Feature's Logs-Odd Estimator

FLOE takes the top r retrieved items (sentences in our case) from the baseline for each query (where r is the number of known relevant sentences for a given query) and, first, computes the probability estimates for this set as follows. Let T be the set of top r relevant sentences and $\bar{T}$ be the remaining sentences in the collection. The estimate for this set is defined as: $\log \frac{p(I|T)}{p(I|\bar{T})} \approx \log \frac{p(I|T)}{p(I)}$.

This value represents how the baseline behaves with respect to the feature I. By subtracting this weight from indep we account for the part of the feature weight that is not captured by the baseline:

$$FLOE(I, R, T) = \log \frac{p(I|R)}{p(I)} - \log \frac{p(I|T)}{p(I)} = \log \frac{p(I|R)}{p(I|T)} \quad (5)$$

The $p(I|T)$ probabilities for tfisf are shown in Figure 1 and Figure 2 represents indep, $\log \frac{p(I|T)}{p(I)}$ and FLOE. The FLOE curve *corrects* the behavior of the baseline to achieve the overall adjustment suggested by indep. The FLOE's adjustment for F_{pos} is erratic (as argued above, we do not expect any benefit from adjustments based on F_{pos}). In contrast, FLOE suggests clearly that the baseline does not retrieve sufficient subjective material in terms of F_{subj}, F_{neg} and F_{opt}.

4. EXPERIMENTS

In our evaluation, we considered the TREC novelty datasets in 2003 and 2004. These test collections supply relevance judgments at sentence level for each topic. The TREC novelty datasets were built as follows. Each collection contains 50 topics. For each topic, a ranked set of documents was obtained by NIST by using a regular retrieval system. Sentence-tagged documents were supplied to participants so that they were asked to find relevant sentences for each topic. Given a topic, each sentence was evaluated as relevant or non-relevant and, therefore, the list of relevance judgments for each topic is complete. The average percentage of relevant sentences in TREC 2003 and TREC 2004 is 38% and 19%, respectively.

Our preprocessing of documents and queries consisted simply on stopword removal. The experiments reported here were run with short queries, constructed from the TREC title field. We planned two different train-test configurations: training with TREC 2003 and testing with TREC 2004, and vice versa. In this way, we can check how robust the opinion-based methods are with respect to the test collection.

Performance was measured in terms of P@10 and MAP (mean average precision). Statistical significance was estimated using the paired t-test (confidence levels of 95% and 99%, marked with $*$ and †, respectively).

The training stage consists of applying FLOE (equation 5) in the training collection to obtain query-independent adjustments such as those shown in Figure 2. Next, these adjustments are applied in the test collection. Given a sentence S and a query Q, the final similarity associated to a sentence is: $tfisf(S, Q) + FLOE(I, R, T_{tfisf})$[3].

Table 1 reports the performance of this method for the test collections. The best results are bolded.

		tfisf+FLOE				
	tfisf (basel.)	F_{subj} (accuracy classifier)	F_{subj} (precision classifier)	F_{neg}	F_{pos}	F_{opt}
test: TREC 2003 (train: TREC 2004)						
P@10	.7480	**.7720***	.7560	.7600	.7500	.7520
$\Delta\%$		(+3.2%)	(+1.0%)	(+1.6%)	(+0.3%)	(+0.5%)
MAP	.3851	**.4012***†	.3908*†	.3907*†	.3829	.3909*†
$\Delta\%$		(+4.2%)	(+1.5%)	(+1.5%)	(−0.6%)	(+1.5%)
test: TREC 2004 (train: TREC 2003)						
P@10	.4300	**.4780***†	.4540*	.4480	.4420	.4460
$\Delta\%$		(+11.2%)	(+5.6%)	(+4.2%)	(+2.8%)	(+3.7%)
MAP	.2358	**.2490***†	.2405*†	.2436*†	.2372	.2414*†
$\Delta\%$		(+5.6%)	(+2.0%)	(+3.3%)	(+0.6%)	(+2.4%)

Table 1: Retrieval performance of tfisf and tfisf+FLOE in the test collections.

The adjustment modelled by FLOE leads to improvements in performance and most of them are statistically significant.

First, the number of positive terms in a sentence, F_{pos}, does not lead to statistical significant improvements. As argued in Section 3.2, we already expected this outcome for F_{pos} because there is not a clear distinction between $p(F_{pos})$

[3]In the following, we use the notation T_{model} to clarify the *model* used to obtain the retrieved set.

and $p(F_{pos}|R)$. Second, the models incorporating the F_{neg} and F_{opt} features outperform clearly the baseline with both performance measures but the improvements are only statistically significant with MAP. Third, F_{subj} appears to be the strongest feature. Detecting subjective sentences with the accuracy-based classifier, which is less stringent than the precision-based classifier, and including this evidence into the SR model leads to very significant improvements in both P@10 and MAP.

4.1 Other functional forms inspired by FLOE

In the previous section we showed that opinion-based features help significantly to retrieve relevant sentences. This positive outcome motivated us to go further and test other functional forms inspired by FLOE. Observe that the adjustments suggested by FLOE (e.g. Figure 2) might be less trustworthy in the regions of the plot with fewer examples. This means that the right-hand end of the F_{neg}, F_{pos} and F_{opt} plots might be misleading. Note also that the forms of the FLOE curves can be easily approximated by simple functions such as lines. These functional forms might generalize better than the original FLOE adjustment and, therefore, they would avoid overfitting. We therefore propose in this section other alternatives to modify the relevance weight with opinion-based evidence. Given a query-independent feature I, we tested the following functions:

$$log(I) = w \cdot \log(I+1) \quad (6)$$

$$linear(I) = w \cdot I \quad (7)$$

$$step(I,w) = \begin{cases} 0 & \text{, if } I = 0 \\ w & \text{, otherwise} \end{cases} \quad (8)$$

where w is a weight that will be tuned in the training stage. The training step for these new adjustments consists simply of tuning w to optimize performance[4]. The test results are shown in Table 2. For F_{subj} we report only the linear function's results because this feature is binary and, therefore, all methods are virtually equivalent.

The relative merits of F_{subj}, F_{neg}, F_{pos} and F_{opt} remain the same: F_{subj} is the strongest feature while F_{pos} is the weakest feature. The new adjustments perform clearly better than the original FLOE's adjustment. Overall, there is no major difference between the functional forms tested but, in terms of statistical significance, the step function looks slightly worse than the others.

The experiments reported demonstrate that opinion-based features are important components that should not be disregarded when retrieving sentences. As a matter of fact, the performance of a state of the art SR model improves very significantly when opinion-based features are included (e.g. F_{subj} leads to 9-26% improvements).

5. CONCLUSIONS

The SR models proposed in this paper (either the ones derived directly from FLOE or those ones inspired by FLOE) outperform significantly a very competitive SR baseline. We provided experimental evidence to show that the subjectivity of a sentence, the number of terms with negative orientation and the number of opinionated terms are sentence features that help to estimate relevance.

The use of opinion-based features in SR is a novel contribution and few attempts have been made in the literature to use opinions for SR [3, 4]. The results reported here opens up a new line of investigation: leveraging different forms of prior information in order to improve baseline retrieval. In this respect, we will study different retrieval scenarios trying to understand when and why subjective content is more amenable to users.

[4]We tested w with values ranging from 0 to 10 in steps of 0.1.

		tfisf+function(I)				
	tfisf (basel.)	F_{subj} (accuracy classifier)	F_{subj} (precision classifier)	F_{neg}	F_{pos}	F_{opt}
test: TREC 2003 (train: TREC 2004)						
P@10 ***log***				.7740	.7740	.7860
Δ%				(+3.5%)	(+3.5%)	(+5.0%)
linear	.7480	**.8300***†	.7900	.7700	.7600*	.7760
Δ%		(+11.0%)	(+5.6%)	(+2.9%)	(+1.6%)	(+3.7%)
step				.7860	.7680	.7720
Δ%				(+5.1%)	(+2.7%)	(+3.2%)
MAP ***log***				.3988*†	.3886*†	.4023*†
Δ%				(+3.6%)	(+0.9%)	(+4.5%)
linear	.3851	**.4213***†	.3986*†	.3965*†	.3886*†	.3982*†
Δ%		(+9.4%)	(+3.5%)	(+3.0%)	(+0.9%)	(+3.4%)
step				.3993*†	.3884*†	.4019*†
Δ%				(+3.7%)	(+0.9%)	(+4.4%)
test: TREC 2004 (train: TREC 2003)						
P@10 ***log***				.4760	.4480	.4680*
Δ%				(+10.7%)	(+4.2%)	(+8.8%)
linear	.4300	**.5420***†	.4920	.4740*	.4460	.4700
Δ%		(+26.1%)	(+14.4%)	(+10.2%)	(+3.7%)	(+9.3%)
step				.4500	.4460	.4560
Δ%				(+4.7%)	(+3.7%)	(+6.1%)
MAP ***log***				.2501*†	.2361	.2517*†
Δ%				(+6.1%)	(+0.1%)	(+6.7%)
linear	.2358	**.2686***†	.2451*†	.2496*†	.2356	.2502*†
Δ%		(+13.9%)	(+3.9%)	(+5.9%)	(−0.1%)	(+6.1%)
step				.2507*†	.2355	.2484*†
Δ%				(+6.3%)	(−0.1%)	(+5.3%)

Table 2: Retrieval performance of tfisf and tfisf+function(I) in the test collections.

Acknowledgements: This research was co-funded by FEDER and *Xunta de Galicia* under projects 07SIN005206PR and 2008/068.

6. REFERENCES

[1] J. Allan, C. Wade, and A. Bolivar. Retrieval and novelty detection at the sentence level. In *Proceedings of ACM SIGIR 2003*, 314–321, Toronto, Canada, 2003.

[2] N. Craswell, S. Robertson, H. Zaragoza, and M. Taylor. Relevance weighting for query independent evidence. In *Proceedings of ACM SIGIR 2005*, pages 416–423, 2005.

[3] S. Kim, D. Ravichandran, and E. Hovy. ISI Novelty Track System for TREC 2004. In *Proceedings of TREC 2004*, 2004.

[4] X. Li, and W.B. Croft. An information pattern based approach to novelty detection. In *Information Processing and Management*, 44(3):1159-1188, 2008.

[5] D. E. Losada. A study of statistical query expansion strategies for sentence retrieval. In *Proceedings of ACM SIGIR 2008 Workshop on Focused Retrieval (Question Answering, Passage Retrieval, Element Retrieval)*, 2008.

[6] D. E. Losada and R. T. Fernández. Highly frequent terms and sentence retrieval. In *Proceedings of SPIRE 2007*, 217–228, Santiago de Chile (Chile), 2007.

[7] B. Pang and L. Lee. Opinion mining and sentiment analysis. In *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, 2008.

[8] B. Pang and L. Lee. Using very simple statistics for review search: An exploration. In *Proceedings of COLING: Companion volume: Posters*, 73–76, Manchester, UK, 2008.

[9] J. Wiebe and E. Riloff. Creating subjective and objective sentence classifiers from unannotated texts. In *Proceedings of CICLing 2005*, 486–497, Mexico, 2005.