A study on evaluation on opinion retrieval systems

Giambattista Amati
Fondazione Ugo Bordoni
Rome, Italy
gba@fub.it

Giuseppe Amodeo
Dept. of Computer Science,
University of L'Aquila
L'Aquila, Italy
gamodeo@fub.it

Valerio Capozio
Dept. of Mathematics,
University of Rome "Tor Vergata"
Rome, Italy
valeriocapozio@gmail.com

Carlo Gaibisso
Istituto di Analisi dei Sistemi ed Informatica "Antonio Ruberti" - CNR
Rome, Italy
carlo.gaibisso@iasi.cnr.it

Giorgio Gambosi
Dept. of Mathematics,
University of Rome "Tor Vergata"
Rome, Italy
gambosi@mat.uniroma2.it

ABSTRACT

We study the evaluation of opinion retrieval systems. Opinion retrieval is a relatively new research area, nevertheless classical evaluation measures, those adopted for ad hoc retrieval, such as MAP, precision at 10 etc., were used to assess the quality of rankings. In this paper we investigate the effectiveness of these standard evaluation measures for topical opinion retrieval. In doing this we split the opinion dimension from the relevance one and use opinion classifiers, with varying accuracy, to analyse how opinion retrieval performance changes by perturbing the outcomes of the opinion classifiers. Classifiers could be studied in two modalities, that is either to re-rank or to filter out directly documents obtained through a first relevance retrieval. In this paper we formally outline both approaches, while for now focussing on the filtering process.

The proposed approach aims to establish the correlation between the accuracy of the classifiers and the performance of the topical opinion retrieval. In this way it will be possible to assess the effectiveness of the opinion component by comparing the effectiveness of the relevance baseline with that of the topical opinion.

Categories and Subject Descriptors

H.3.0 [**Information Storage and Retrieval**]: General; H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Theory, Experimentation

Keywords

Sentiment Analysis, Opinion Retrieval, Opinion Finding, Classification

Appears in the Proceedings of the 1st Italian Information Retrieval Workshop (IIR'10), January 27–28, 2010, Padova, Italy.
http://ims.dei.unipd.it/websites/iir10/index.html

1. INTRODUCTION

Sentiment analysis aims to documents classification, according to opinions, sentiments, or, more generally, subjective features contained in text. The study and evaluation of efficient solutions to detect sentiments in text is a popular research area, and different techniques have been applied coming from natural language processing, computational linguistics, machine learning, information retrieval and text mining.

The application of sentimental analysis to Information Retrieval goes back to the novelty track of TREC 2003 [13]. Topical opinion retrieval is also known as *opinion retrieval* or *opinion finding* [4, 9, 11]. In [5, 3, 2, **?**] dictionary-based methodologies for topical opinion retrieval are proposed. An application of opinion finding to blogs was introduced in the Blog Track of TREC 2006 [8]. However, there is not yet a comprehensive study of evaluation of topical opinion systems, and in particular of the interaction and correlation between relevance and sentiment assessments.

At first glance, evaluation of opinion retrieval systems seems to not deserve any further investigation or extra effort with respect to the evaluation of conventional retrieval systems. Traditional evaluation measures, such as the Mean Average Precision (MAP) or the precision at 10 [8, 6, 10, 11], can be still used to evaluate rankings of opinionated documents that are also assessed to be relevant to a given topic. However, if we give a deeper look at the performance of topical opinion systems we are struck by the diversity in the observed values of performance. For example the best run for topic relevance in the blog track of TREC 2008 [10] achieves a MAP value equal to 0.4954, that drops to 0.4052, as concerns the MAP of opinion, in the opinion finding task. Performance degradation is as expected because any variable which is additional to relevance, i.e. the opinion one, must deteriorate the system performance. However, we do not have yet a way to set apart the effectiveness of the opinion detection component and evaluate how effective it is, or to determine whether and to which extent, the relevance and opinion detection components are influenced by each other. It seems evident that an evaluation methodology or at least some benchmarks are needed to make it possible to assess how effective the opinion component is. To exemplify: how effective is the performance value of opinion MAP 0.4052 when we start from an initial relevance MAP of 0.4954? It is indeed a matter of fact that opinion MAP in TREC [8, 6, 10], seems to be highly dependent on the relevance MAP of the first-pass retrieval [9].

The general issue is thus the following: can we assume that absolute values of MAP can be used as they are to compare different tasks, in our case the topical opinion and the ad hoc relevance task; and thus: evaluation measures can be used without any MAP normalization to compare or to assess the state of the art of different techniques on opinion finding?

At this aim, we introduce a completely novel methodological framework which:

- provides a bound for the best achievable opinion MAP, for a given relevance document ranking;
- predicts the performance of topical opinion retrieval given the performance of the topic retrieval and opinion detection;
- viceversa, provides whether a given opinion detection technique gives a significant or marginal contribution to the state of the art;
- investigates the robustness of evaluation measures for opinion retrieval effectiveness.
- indicates what re-ranking or filtering strategy is best suited to improve topical retrieval by opinion classifiers.

This paper is organized as follows. The proposed evaluation method is presented in sections 2 and 4; section 3 introduces the collection used for tests. Results are presented in section 5, and conclusions follow in section 6.

2. EVALUATION APPROACH

An opinion retrieval system is based on a topic retrieval and an opinion detection subsystem [9]: different kinds of "information" are retrieved and weighted in order to generate a final ranking of documents that reflects their relevance with both topic and opinion content. To analyse the effectiveness of the whole system, we should be able to quantify not only the performance of the final result, but also the contribution of each subsystem. As usual, the evaluation metric used in literature for the final ranking is the MAP. But MAP (of relevance and opinion) for the final ranking is not sufficient to fully assess the performance of the whole system: the contribution of each component, taken separately, needs to be identified.

The input to the proposed topical opinion evaluation process is the relevance *baseline*, i.e. the ranking of documents generated by the topic retrieval system, here considered as a black box. The effectiveness of the topic retrieval component is measured by the MAP of opinion and relevance of this baseline.

The evaluation of the effectiveness of the opinion detection component, relies on artificially defined classifiers of opinion. The artificial classifier $\mathcal{C}^k_{\mathcal{O}}$ classifies documents as opinionated, $\mathcal{O}$, or not opinionated, $\overline{\mathcal{O}}$, with accuracy k, $0 \leq k \leq 1$. The classification process is independent from the topic relevance of documents. To achieve accuracy k $\mathcal{C}^k_{\mathcal{O}}$ properly classifies each document with probability k.

Therefore the number of misclassified documents is $(1-k)\cdot n$, where n is the number of classified documents. Assuming the independence between opinion and relevance, the misclassified documents will be distributed randomly between relevant and not relevant.

The outcomes of these artificial classifiers are then used to modify the baseline. This can be done following two different approaches:

- a *filtering process*: when documents of the baseline are deemed as not opinionated by the classifier, they are removed from the ranking;
- a *re-ranking process*: when documents of the baseline are considered as opinionated by the classifier, they receive a "reward" in their rank.

The filtering process uses the classifier in its classical meaning. This process is particularly suitable to analyse the effectiveness of the technique itself to opinion detection, as a classification task [12], and its effects on topical opinion performance. Opinion filtering also gives some interesting clues on what is the optimal performance achievable by an opinion retrieval technique based on filtering, and also whether filtering strategy is in general superior or not to even very simple re-ranking strategies.

In the re-ranking process a "reward" function for the documents has to be defined. In such a case we introduce bias in assigning correct rewards, and we thus may observe the effectiveness of a re-ranking algorithm as long as the opinion detection performance changes.

By "comparing" the results of an opinion retrieval system with the filtering process, or the re-ranking process at several levels of accuracy, we can obtain relevant clues about:

- the overall contribution introduced by the opinion system only and its robustness;
- the effectiveness of the opinion detection component;

In the following we formally describe both the approaches and focus on the experimentation concerning the filtering process only.

3. EXPERIMENTATION ENVIRONMENT

We used the BLOG06 [7] collection and the data sets of the Blog Track of TREC 2006, 2007 and 2008 [8, 6, 10] for our experimentation. Since 2006, Blog Track has an evaluation track on blogs where the main task is opinion retrieval, that is the task of selecting the opinionated blog posts relevant to a given topic [9]. BLOG06 collection size is 148 GB and contains spam as well as possibly non-blogs and non-English pages.

The data set consists of 150 topics and a list, the *Qrels*, in which the relevance and content of opinion of documents are assessed with respect to each topic. An item in the list identifies a topic t, a document d and a judgement of relevance/opinion assigned as follows:

- 0 if d is not relevant with respect to t;
- 1 if d is relevant to t, but does not contain comments on t;
- 2 if d is relevant to t and contains positive comments on t;

- 3 if d is relevant to t and contains neutral comments on t;
- 4 if d is relevant to t and contains negative comments on t.

Note that not relevant documents are not classified according to their opinion content.

In the following, $[x]$ denotes the set of documents labelled by an $x = 0, 1, 2, 3, 4$, and not labelled documents belong to [0] by default.

TREC organizers also provide the best five *baselines*, produced by some participants, denoted by BL_1, $BL_2, \ldots, BL_5$.

4. EVALUATION FRAMEWORK

The behaviour of artificial classifier $\mathcal{C}_{\mathcal{O}}^{k}$ is defined through the *Qrels*. $\mathcal{C}_{\mathcal{O}}^{k}$ predicts the right opinion orientation of each document in the collection by searching it in the *Qrels*. The accuracy k is simulated by the introduction of a bias in the classification. Documents not appearing or assessed as not relevant in the *Qrels*, will be classified according to the distribution of probability of opinionated and not opinionated documents among the relevant ones. Taking into account both relevance and opinion in the test collection we obtain the contingency Table 1. As shown in table 1, the *Qrels* does not provide the opinion classes for not relevant documents. The missing data complicate a little bit, but not much, the construction of our classifiers. To overcome the problem, we assume that

$$Pr(\mathcal{O}|\mathcal{R}) = Pr(\mathcal{O}|\overline{\mathcal{R}}) \quad (1)$$

Equation 1 asserts that there is not a sufficient reason to have a different distribution of opinion among relevant and not relevant documents. An a priori probability, $\Pr(\mathcal{O})$, for opinionated documents is still unknown. However equation 1 implies that $\mathcal{O}$ and $\mathcal{R}$ are independent, thus

$$Pr(\mathcal{O}|\mathcal{R}) = Pr(\mathcal{O}) \quad (2)$$

From equations 1 and 2 follows that

$$Pr(\overline{\mathcal{O}}|\mathcal{R}) = Pr(\overline{\mathcal{O}}|\overline{\mathcal{R}}) = Pr(\overline{\mathcal{O}}) = 1 - Pr(\mathcal{O}) \quad (3)$$

Equations 2 and 3 are equivalent to assume that the set $\{[2] \cup [3] \cup [4]\}$, as defined in Table 1, is a sample of the set of opinionated documents. Thus, without loss of generality, we can define $\Pr(O)$ using only the documents classified as relevant by the *Qrels* as follows:

$$P(\mathcal{O}) = \frac{|\{[2] \cup [3] \cup [4]\}|}{|\{[1] \cup [2] \cup [3] \cup [4]\}|} \quad (4)$$

and consequently

$$P(\overline{\mathcal{O}}) = 1 - P(\mathcal{O}) = \frac{|[1]|}{|\{[1] \cup [2] \cup [3] \cup [4]\}|} \quad (5)$$

In the following we study whether and how the set of relevant and not relevant documents classified as opinionated affects the topical opinion ranking.

We have to say that for both approaches, filtering or reranking, a misclassification may have controversial effects on the effectiveness of the final ranking. If we filter documents by opinions with a classifier, for example, the misclassified and removed not relevant documents may bring a positive contribution to the precision measures, because all opinionated and relevant documents that were below them, will have a higher rank after their removal. Even with the re-ranking approach we have a similar situation, but this precision boosting phenomenon is attenuated by the fact that re-ranking is not based on as drastic decision as that of a removal, and the repositioning of a document does not propagate to all documents that are below it in the original ranking.

	$\mathcal{O}$	$\overline{\mathcal{O}}$
$\mathcal{R}$	$\|\{[2]\cup[3]\cup[4]\}\|$	$\|[1]\|$
$\overline{\mathcal{R}}$	NA	NA

Table 1: the contingency table for an opinion-only classifier for documents in the BLOG06 collection. $\mathcal{R}$ denotes relevance, $\overline{\mathcal{R}}$ non-relevance; $\mathcal{O}$ denotes opinion, $\overline{\mathcal{O}}$ non-opinion. With the notation [x] we refer to the class of documents labelled by $x = 1, 2, 3, 4$ in the *Qrels*.

Together with $\mathcal{C}_{\mathcal{O}}^{k}$, we introduce a random classifier $\mathcal{C}_{\mathcal{O}}^{RC}$ that classifies documents according to the a priori distribution of opinionated documents in the collection. It represents a good approximation of the random behaviour of a classifier. More precisely, this classifier assesses a document as opinionated with probability $P(\mathcal{O})$ and as not opinionated with probability $\Pr(\overline{\mathcal{O}}) = 1 - \Pr(\mathcal{O})$.

4.1 Filtering approach

As already stated, in the filtering approach documents classified as not opinionated are removed from the baseline. Note that while relevant documents contribute and improve the evaluation measure, if correctly classified, the not relevant ones do not contribute directly to this measure.

In conclusion if a not relevant document is classified as opinionated not being actually opinionated, then this misclassification will not affect the evaluation measure. Differently the removal of not relevant documents regardless of their real opinion orientation, always positively affects the ranking, even if misclassified.

For relevant documents instead the misclassification always negatively affects the ranking.

With this approach we can observe how hard is to overcome the baseline, i.e. we can identify how effective must be the opinion detection technique to improve the starting topic retrieval.

4.2 Re-ranking approach

Re-ranking techniques essentially are fusion models [9] that combine a relevance score $s_{\mathcal{R}}(d)$ and an opinion score $s_{\mathcal{O}}(d)$ (or two ranks derived from these scores) for a document d. The new score $s_{\mathcal{OR}}(d)$ is a function of the two non negative scores, $s_{\mathcal{R}}(d)$ and $s_{\mathcal{O}}(d)$:

$$s_{\mathcal{OR}}(d) = f(s_{\mathcal{R}}(d), s_{\mathcal{O}}(d)) \quad (6)$$

Given a classifier $\mathcal{C}_{\mathcal{O}}^{k}$, we define a new score $s_{\mathcal{OR}}^{\mathcal{C}}(d)$ based on the outcomes of $\mathcal{C}_{\mathcal{O}}^{k}$ according to which the baseline is re-ranked. $s_{\mathcal{OR}}^{\mathcal{C}}(d)$ is defined as follows:

$$s_{\mathcal{OR}}^{\mathcal{C}}(d) = \begin{cases} f(s_{\mathcal{R}}(d), s_{\mathcal{O}}(d)) & \text{if } d \in_{\mathcal{C}_{\mathcal{O}}^{k}} O \\ f(s_{\mathcal{R}}(d), 0) & \text{if } d \notin_{\mathcal{C}_{\mathcal{O}}^{k}} O \end{cases} \quad (7)$$

where $\in_{\mathcal{C}_{\mathcal{O}}^{k}}$ denotes the classifier outcome, that is when the document is assigned to a given class. Note when $k = 100\%$

and assuming that $f(\cdot,\cdot)$ is a not decreasing function of $s_{\mathcal{O}}(\cdot)$, i.e. $f(s_{\mathcal{R}}(d),x) \geq f(s_{\mathcal{R}}(d),x')$, $\forall x \geq x'$, the opinion MAP of any ranking based on $s_{\mathcal{OR}}(\cdot)$ does not exceed that based on $s^{\mathcal{C}}_{\mathcal{OR}}(\cdot)$.

All the above considerations can be further extended to the case in witch the $s_{\mathcal{OR}}(d)$ is based on the ranks of d instead of on its scores (of relevance and opinion).

5. EXPERIMENTATION RESULTS

In this paper we report the experimentation results for the filtering approach. The filtering process has been repeated 20 times for each baseline and for accuracy $k = 0.5$, 0.6, 0.7,0.8,0.9,1. Mean values of the MAPs are reported.

Table 2 reports, in decreasing order, the relevance MAPs (MAP_R) and the opinion MAPs (MAP_O) for each baseline.

Baselines		
	MAP_R	MAP_O
BL_4	0.4776	0.3542
BL_5	0.4247	0.2974
BL_3	0.4079	0.3007
BL_1	0.3540	0.2470
BL_2	0.3382	0.2657

Table 2: MAP of relevance (MAP_R) and opinion (MAP_O) of the five baselines.

In figure 1 MAP values are reported for each baseline as long as the accuracy of classifiers changes. The dotted lines represent the baselines opinion MAPs and the dot-dashed lines represent the baseline relevance MAPs. The MAP values of random classifier is also reported as the dashed lines in the graphs.

Analysing the MAP trend we can infer the following observations:

1. the baseline MAP_R is an upper bound for the MAP_0 obtained with a filtering approach;

2. the random classifier always deteriorate the performance of the baseline MAP_0.

3. the minimal accuracy needed to improve by filtering the baseline MAP_0 is very high, at least 80%;

4. there is a linear correlation between the MAP_0 achievable by a classifier with accuracy k and the accuracy itself.

First three remarks says that filtering strategy is very dangerous for MAP_0 performance, that is removing documents affects greatly the performance of the topical opinion retrieval.

From the above considerations, we may conclude that the opinion retrieval task is not easy and that having good results with a filtering approach requires a too high accuracy. The experimentation instead allows us to identify a plausible range for the MAP achievable by an opinion retrieval system: the classifier with accuracy 100% and the random classifiers obtains performance that can be considered as thresholds for the best and the worst opinion detection system. It is also evident that higher the baseline MAP is, higher the accuracy of classifier must be to introduce some benefits with a filtering approach with respect to relevance only retrieval.

6. CONCLUSIONS AND FUTURE WORKS

The opinion retrieval problem seems to be a relatively hard task: the combination of two variables like topic relevance and opinion, requires a deep analysis on their correlation. From the results of TREC competitions [8, 6, 10, 9], emerges the lack of exhaustive evaluations measures: the MAP, Precision at 10 and R-Precision are not sufficient alone to give a complete analysis on the systems performances.

Up to now we have studied only the filtering of documents by opinions. This strategy however requires a very high accuracy of the classification. We will compute the study with re-ranking approach starting from the approach used in [1, 2].

Our approach is able to provide an indicative accuracy of the opinion component of the topical opinion retrieval system. It also allows us to propose an evaluation framework, able to evaluate the effectiveness of opinion retrieval systems.

7. REFERENCES

[1] 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 *Proc. of the 16th Text Retrieval Conference (TREC)*, 2007.

[2] G. Amati, G. Amodeo, M. Bianchi, C. Gaibisso, and G. Gambosi. *A uniform theoretic approach to opinion and information retrieval*, in *Intelligent Information Access*, G. Armano, M. de Gemmis, G. Semeraro, and E. Vargiu (eds.) Studies in Computational Intelligence. Springer, to appear.

[3] J. Skomorowski and O. Vechtomova. Ad hoc retrieval of documents with topical opinion. In G. Amati, C. Carpineto, and G. Romano, editors, *ECIR*, volume 4425 of *Lecture Notes in Computer Science*, pages 405–417. Springer, 2007.

[4] K. Eguchi and V. Lavrenko. Sentiment retrieval using generative models. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 345–354, Sydney, Australia, July 2006. Association for Computational Linguistics.

[5] G. Mishne. Multiple ranking strategies for opinion retrieval in blogs. In *The Fifteenth Text REtrieval Conference (TREC 2006) Proceedings*, 2006.

[6] C. Macdonald, I. Ounis, and I. Soboroff. Overview of the trec-2007 blog track. In *Proc. of the 16th Text Retrieval Conference (TREC)*, 2007.

[7] Crag Macdonald and Iadh Ounis. The trec blogs06 collection : Creating and analysing a blog test collection. Technical report, University of Glasgow Scotland, UK, 2006.

[8] I. Ounis, M. de Rijke, C. Macdonald, G. A. Mishne, and I. Soboroff. Overview of the trec-2006 blog track. In *TREC 2006 Working Notes*, 2006.

[9] I. Ounis, C. Macdonald, and I. Soboroff. On the trec blog track. In *Proc. of the 2nd International Conference on Weblogs and Social Media (ICWSM)*, 2008.

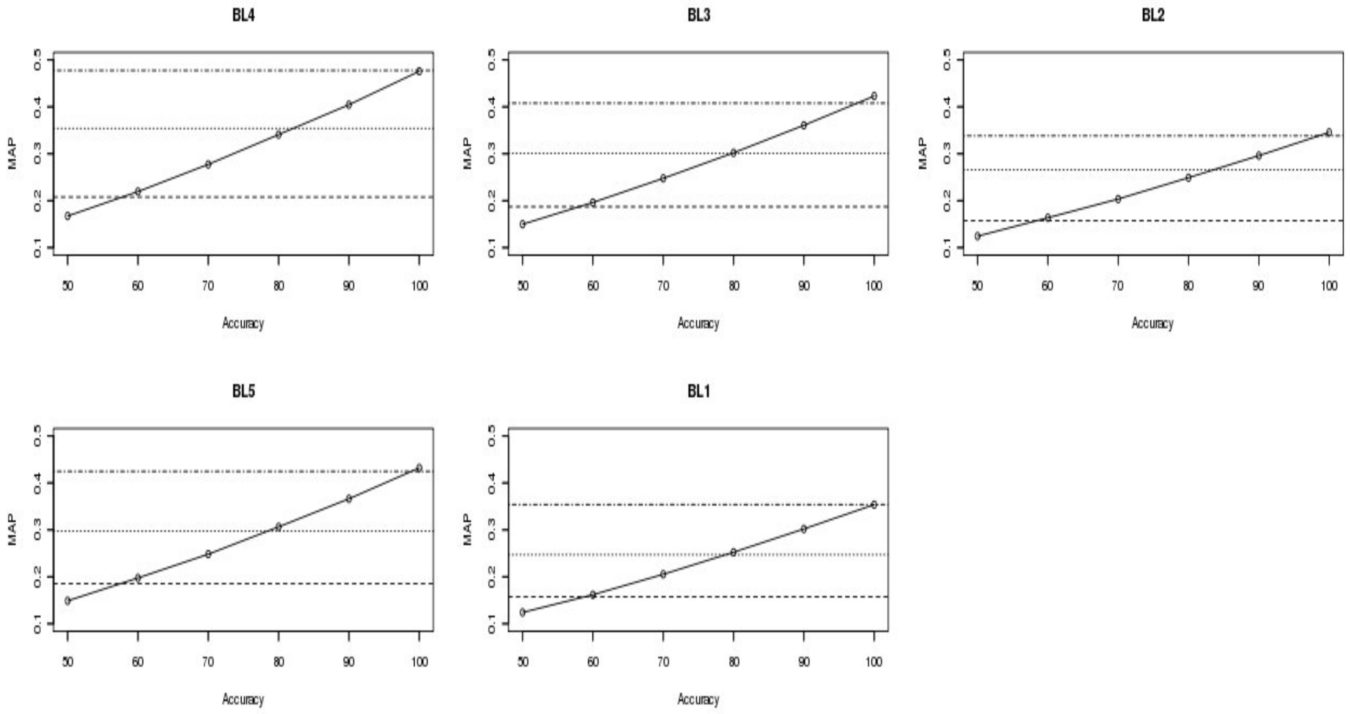

Figure 1: MAPs of opinion of the baselines filtered by $\mathcal{C}_{\mathcal{O}}^{k}$ and for $k = 0.5, 0.6, 0.7, 0.8, 0.9, 1$. The opinion MAPs (dotted lines) and relevance MAPs (dot-dashed lines) of the baselines are also reported. Finally dashed lines show the opinion MAPs for the baselines filtered by $\mathcal{C}_{\mathcal{O}}^{RC}$.

[10] I. Ounis, C. Macdonald, and I. Soboroff. Overview of the trec-2008 blog track. In *Proc. of the 17th Text Retrieval Conference (TREC)*, 2008.

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

[12] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. In *Proc. of the ACL-02 conference on Empirical Methods in Natural Language Processing*, pages 79–86, 2002.

[13] Ian Soboroff and Donna Harman. Overview of the trec 2003 novelty track. In *TREC*, pages 38–53, 2003.