Novelty Detection Using Local Context Analysis

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

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

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval Models, Information Filtering

General Terms

Experimentation

Keywords

Local Context Analysis, Novelty Detection

1. INTRODUCTION

The aim of this work is to determine the utility of Local Context Analysis (LCA)[5] for retrieval of relevant and novel sentences. LCA has been successful in different areas and we check here whether this method is also useful to drive the selection of novel material. We adopt the Novelty task as defined in the TREC conference [2, 4, 3]. Giving a set of documents associated to a topic, the task consists of finding the relevant and novel sentences. This problem is interesting for many areas, such as text summarization, web information access, question answering, etc. Some researchers have proposed that the estimation of novelty for a given sentence should be based on the set of seen sentences that share common meanings [6]. In this way, the degree of redundancy of a sentence s_i is not influenced by past sentences that are totally unrelated to s_i. The intuition is that novelty estimation might be more robust if focused on this set of terms. In our work we pursue a similar idea because we apply LCA to focus the estimation of novelty on query-related terms.

2. THE NOVELTY TASK AND LCA

The groups participating in the Novelty task start from a common ranking of documents for each query. Two different subtask are proposed: 1) to produce a ranking of relevant sentences and 2) to filter out redundant sentences from this ranking. Successful algorithms tested for this task apply usually some popular IR model to rank sentences given a query (e.g. variants of tf-idf applied at the sentence level [1]). Next, in order to estimate how redundant the sentences are, some methods have been proposed to compute the overlapping between each sentence and the previously seen sentences. We have chosen two baseline methods which are simple and robust [1]: NewWords and SetDif. NewWords counts the number of words in the current sentence, s_i, which did not occur in the previously seen sentences:

$$N_{nw}(s_i|s_1,...,s_{i-1}) = \left| W_{s_i} \cap \overline{\bigcup_{j=1}^{i-1} W_{s_j}} \right|$$

where W_{s_i} is the set of words in the sentence s_i.

SetDif computes the number of different words between each sentence s_i and the previously seen sentence that is the most similar to s_i:

$$N_{sd}(s_i|s_1,...,s_{i-1}) = \min_{1 \leqslant j \leqslant i-1} N_{sd}(s_i|s_j)$$

$$N_{sd}(s_i|s_j) = \left| W_{s_i} \cap \overline{W_{s_j}} \right|$$

We propose variants of these methods to estimate the novelty score focusing on query-related terms. We expect to improve performance when the novelty scores consider only terms that are highly related to the query.

2.1 LCA

LCA is a method based on the idea that a common term from the top-ranked relevant documents (or passages) will tend to co-occur with query terms within the top-ranked documents (or passages) [5]. We apply LCA to produce a set of query-related terms and the novelty scores are adjusted accordingly. The importance of the terms in the top ranked sentences is computed as:

$$bel(q,t) = \prod_{t_i \in q} \delta + log(af(t,t_i))idf_t/log(n))^{idf_i}$$

where t is a term, N is the number of sentences in the collection, N_i is the number of sentences containing the term t_i, ft_{ij} is the number of occurrences of the term t_i in the sentence p_j, ft_j is the number of occurrences of the term t in th sentence p_j, $idf_i = \min(1.0, log_{10}(N/N_i)/5.0)$, $af(t,t_i) = \sum_{j=1}^{n} ft_{ij} ft_j$, and δ is 0.1 (a constant) to avoid zero bel value.

This measure can be applied to rank terms in decreasing order of estimated importance given a query. Selecting the

SIGIR'07, July 23–27, 2007, Amsterdam, The Netherlands.
Copyright 2007 ACM 978-1-59593-597-7/07/0007 ...$5.00.

		NW	**NW LCA**			
			10 t.	**50 t.**	**100 t.**	**all t.**
T2002	***P@5***	0.200	0.204	0.229	0.245	0.237
	P@10	0.180	0.151	0.190	0.222*	0.235*
T2003	***P@5***	0.596	0.532	0.552	0.572	0.596
	P@10	0.572	0.478*	0.538	0.562	0.580
T2004	***P@5***	0.224	0.248	0.288*	0.284*	0.256
	P@10	0.252	0.190*	0.246	0.264	0.274

Table 1: NewWords vs. NewWords with LCA

		SD	**SD LCA**			
			10 t.	**50 t.**	**100 t.**	**all t.**
T2002	***P@5***	0.208	0.216	0.220	0.241	0.233
	P@10	0.184	0.188	0.214	0.229	0.233*
T2003	***P@5***	0.568	0.564	0.540	0.564	0.584
	P@10	0.580	0.536	0.544	0.558	0.590
T2004	***P@5***	0.236	0.256	0.296	0.308*	0.264
	P@10	0.256	0.220	0.262	0.272	0.286

Table 2: SetDif vs. SetDif with LCA

top ranked terms we can conform a query-oriented vocabulary (T_q). Using this vocabulary, we compute NewWords and SetDif for each sentence as follows:

$$N_{LCA_{nw}}(s_i|s_1,...,s_{i-1}) = \left| W_{LCA_{s_i,q}} \cap \overline{\bigcup_{j=1}^{i-1} W_{LCA_{s_j,q}}} \right|$$

and

$$N_{LCA_{sd}}(s_i|s_1,...,s_{i-1}) = \min_{1 \leqslant j \leqslant i-1} N_{LCA_{sd}}(s_i|s_j)$$

$$N_{LCA_{sd}}(s_i|s_j) = \left| W_{LCA_{s_i}} \cap \overline{W_{LCA_{s_j}}} \right|$$

where $W_{LCA_{s_i,q}} = W_{s_i} \cap T_q$.

3. EXPERIMENTS

We used the three different collections of data which were made available in the context of the TREC Novelty tracks in 2002, 2003 and 2004 [2, 4, 3]. In 2002 and 2003, the ranking of documents provided by NIST consists only of relevant documents. In 2004, the collection is more realistic because the ranks of documents contain relevant and irrelevant material.

To generate an initial rank of sentences we applied a variation of tf-idf which proved successful in the past [1]. Given these ranks, the top 25 ranked sentences[1] are mined for selecting important terms using LCA. This gives us the query-oriented vocabulary T_q and, subsequently, sentences are re-ranked using $N_{LCA_{nw}}$ and $N_{LCA_{sd}}$. The top 10% sentences of this ranking are used for evaluation. We made experiments with varying sizes of this vocabulary to check the stability of the method. The evaluation measures applied are precision at 5 (P@5) and precision at 10 sentences (P@10).

In Table 1 we show the performance values using NewWords and NewWords with LCA applying different vocabulary sizes (10, 50, 100 and all terms in the top 25 ranked sentences). Analogously, in Table 2 we report results for the SetDif method. Results indicated with a star are statistically significant using a t-test at the $p < .05$ level.

In 2003, the baseline performs very well because of the high population of relevant sentences in the collection [4]. Hence, it is very difficult to improve the results because any reasonable sentence retrieval strategy yields a good top 10. In the other two collections the application of LCA yielded significant improvements.

The results indicate that the larger the vocabulary is the better the precision is. With 10 terms the method does not estimate redundance satisfactorily because all the decisions are made based on very few terms. On the other hand, if vocabularies contain all terms in the top 25 sentences then redundance is estimated successfully.

LCA seems useful in terms of P@5 but its utility is questionable if the aim is to retrieve 10 good sentences. In such case, selecting simply all terms in the top 25 sentences is the most robust approach. To the best of our knowledge, this sort of vocabulary selection, which is a form of pseudo-relevance feedback for novelty purposes, has not been applied in the literature.

4. CONCLUSIONS

We have presented the results of our attempts to identify relevant and novel sentences in a ranked list of documents using different methods and their variants using LCA.

Although NewWords and SetDif are competitive methods for novelty detection, our results indicate that precision at top ranks might be further improved if redundancy decisions are made in terms of a more focused vocabulary. Nevertheless, it is still unclear whether such vocabulary should be selected using LCA. Given our current results, a simple method (based on extracting the terms appearing in the top 25 sentences) performs well and does not require LCA. Anyway, in the future we will keep studying the effects of the vocabulary size on novelty detection.

5. ACKNOWLEDGEMENTS

This work was partially supported by national project TIN2005-08521-C02-01 and Galician network 2006/23. David E. Losada belongs to the “Ramón y Cajal” program, whose funds come from MEC and the FEDER program.

6. REFERENCES

[1] J. Allan, C. Wade, and A. Bolivar. Retrieval and novelty detection at the sentence level. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2003)*, pages 314–321, 2003.

[2] D. Harman. Overview of the TREC 2002 Novelty Track. In *Proceedings of the eleventh Text REtrieval Conference (TREC 2002)*, 2002.

[3] I. Soboroff. Overview of the TREC 2004 Novelty Track. In *Proceedings of the 13th Text REtrieval Conference (TREC 2004)*, 2004.

[4] I. Soboroff and D. Harman. Overview of the TREC 2003 Novelty Track. In *Proceedings of the twelfth Text REtrieval Conference (TREC 2003)*, 2003.

[5] J. Xu and W. B. Croft. Improving the effectiveness of information retrieval with local context analysis. *ACM Transactions on Information Systems*, 18(1):79–112, 2000.

[6] L. Zhao, M. Zhang, and S. Ma. The nature of novelty detection. *Inf. Retr.*, 9(5):521–541, 2006.

[1] Preliminary experiments showed that 25 sentences is a reasonable number for estimating the query-oriented vocabulary.