Novelty Detection Using Local Context Analysis

Ronald T. Fernández **David E. Losada**
ronald.teijeira@rai.usc.es dlosada@usc.es
Departamento de Electrónica y Computación
Universidad de Santiago de Compostela,SPAIN

Novelty Detection:

Go beyond the traditional relevance-oriented ranking of documents.

Filter redundant material ⟶ increase user satisfaction.

Interesting subject in many areas: text summarization, web information access, question answering, etc.

TREC Novelty Tracks:

Find relevant and novel sentences in a ranked set of documents (constructed from a query).

Current methods to detect novelty (e.g. NewWords [1]) are based on word counts and overlapping measures with the previously seen sentences.

Problem: terms unrelated to the query can trigger novelty.

Motivation:

Aim: Determine the utility of Local Context Analysis (LCA) for retrieval of relevant and novel sentences.

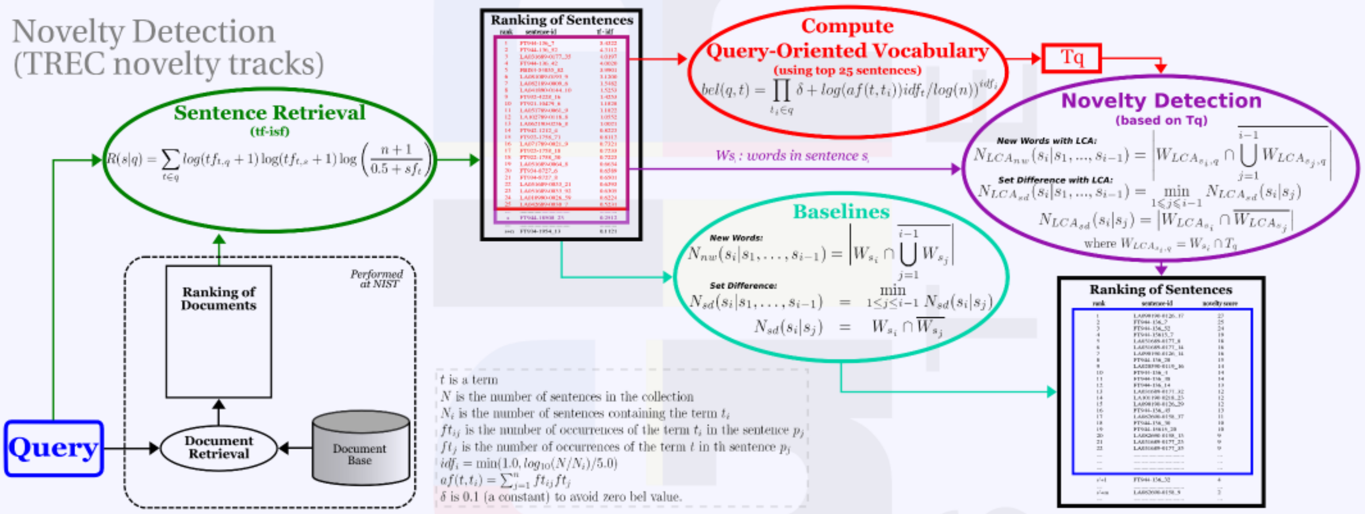
LCA: A common term from the top-ranked relevant documents will tend to co-occur with query terms within the top ranked documents.

Effective method to estimate the importance of terms (e.g. for QE).

Focus the novelty detection on a **vocabulary related to the query**.

Is LCA useful to drive novelty detection?

Novelty Detection (TREC novelty tracks)



$$R(s|q) = \sum_{t \in q} log(tf_{t,q}+1)\log(tf_{t,s}+1)\log\left(\frac{n+1}{0.5+sf_t}\right)$$

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

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

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

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

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

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

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

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

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
$idf_i = \min(1.0, log_{10}(N/N_i)/5.0)$
$af(t,t_i) = \sum_{j=1}^{n} ft_{ij} ft_j$
δ is 0.1 (a constant) to avoid zero bel value.

Experiments

TREC 2002, 2003 and 2004 novelty tracks' data.

Baselines: NewWords and SetDif [1].

Select the top 25-retrieved sentences to build the vocabulary (Tq).

Experiments with varying size of the vocabulary Tq.

Results:

TREC 2003: many relevant sentences ⟶ no improvements (at least, in terms of P@5).

TREC 2002, 2004: harder collections ⟶ LCA more useful.

Taking a large number of terms in the top 25 sentences is the best choice. Larger vocabulary ⟶ better precision.

LCA looks promising to enhance the retrieval of a few novel sentences.

	NewWords	NewWords LCA			
		10 terms	*50 terms*	*100 terms*	*all terms*
T2002	*0.200*	*0.204*	*0.229*	*0.245*	*0.237*
T2003	*0.596*	*0.532*	*0.552*	*0.572*	*0.596*
T2004	*0.224*	*0.248*	*0.288*	*0.284*	*0.256*

	SetDif	SetDif LCA			
		10 terms	*50 terms*	*100 terms*	*all terms*
T2002	*0.208*	*0.216*	*0.220*	*0.241*	*0.233*
T2003	*0.568*	*0.564*	*0.540*	*0.564*	*0.584*
T2004	*0.236*	*0.256*	*0.296*	*0.308*	*0.264*

P@5 results for
TREC 2002, 2003 and 2004
and different sizes of vocabulary

References:

[1] J. Allan, C. Wade, 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, pages 314-321, 2003

[2] D. Harman. Overview of the TREC 2002 Novelty Track. In Proceedings of the 11th Text REtrieval Conference, 2002

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

[4] I. Soboroff, D. Harman. Overview of the TREC 2003 Novelty Track. In Proceedings of the 12th Text REtrieval Conference, 2003

[5] J. Xu, 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, S. Ma. The nature of novelty detection. Information Retrieval, 9(5): 521-541, 2006