TREC NOVELTY TRACK AT IRIT – SIG

Taoufiq Dkaki[1,2], Josiane Mothe[1,3]

(1) Institut de Recherche en Informatique de Toulouse, 118 Rte de Narbonne, 31062 Toulouse CEDEX
(2) Université Toulouse le Mirail, ISYCOM/GRIMM,
(3) Institut Universitaire de Formation des Maîtres Midi-Pyrénées

Abstract

In TREC 2003, IRIT improved the strategy that was introduced in TREC 2002. A sentence is considered as relevant if it matches the topic with a certain level of coverage. This coverage depends on the category of terms used in the texts. Different types of terms have been defined: highly relevant, scarcely relevant, non-relevant and highly non-relevant. With regard to the *novelty part*, a sentence is considered as novel if its similarity with previously processed sentences and with the n-best-matching sentences does not exceed certain thresholds.

1 Introduction

«The TREC novelty track is designed to investigate systems' abilities to locate relevant and new information within the ranked set of documents retrieved in answer to a TREC topic » [trec.nist.gov].

Retrieving relevant texts is traditionally based on computing a similarity between the representations of the information need (or topic) and the texts. This general statement has been applied to full documents as well as chunks of texts (passage retrieval). Intuitively, the same idea can be applied when sentences retrieval is involved. In TREC 2002 IRIT developed a new strategy in order to detect the relevant sentences. This approach has not been used in a more general context of document retrieval but we did used it previously and partially in document categorization (Mothe, 2002). In our approach a sentence is considered as *relevant* if it matches the topic with a certain level of coverage. This level of coverage depends on the category of the terms used in the texts. Three types of terms were defined for TREC 2002: highly relevant, scarcely relevant and non-relevant. In TREC 2003 we introduced a new class of terms: highly non-relevant terms. Terms from this category are extracted from the narrative parts of the queries that describe what is a non-relevant document. A negative weight can be assigned to these words. With regard to the *novelty part*, a sentence is considered as novel if its similarity with each previously processed and -selected as novel- sentences does not exceed a certain threshold. In addition, this sentence should not be too similar to a virtual sentence made of the n-best-matching sentences.

The results we obtained in TREC 2002 were quite good regarding the 'relevant' subtask. Indeed, for 36 topics (73%), the R*P was higher or equal to the average of the 42 runs which were submitted. In TREC 2003, we improved these results as we obtained 46 topics (92%) for which the F-measure (2*R*P/(R+P)) was equal or higher to the average of the 55 runs submitted. With regard to the 'novelty' part, when considering the retrieved sentences, we also obtained 46 topics (92%) for which the F-measure is higher or equal to the average of the 55 runs. However, an interesting result is that our method is better when there is some noise in the sentence set. Indeed the results are better when considering the retrieved sentences than when considering the relevance sentences only, relatively to other participants' methods (i.e. our system ranks better over the submitted runs). We obtained 41 topics (82%) for which the F-measure is higher or equal to the average of the 55 runs.

This paper is organized as follows: in section 2 we describe the method we used, including the way documents and topics are represented and the strategies we developed for the two sub-tasks (relevant part and novelty part either considering only relevant sentences or all retrieved sentences). In section 3 we present the results and comment them. We also present results we obtained on runs that were not submitted.

2 Description of the method

2.1 Document and topic representation

In our method, topics and sentences are considered as chunks of text. Each chunk is pre-processed the same way in order to extract representative terms. Then, terms extracted from a given topic are categorized into different groups: highly relevant terms (HT), scarcely relevant terms (LT) and highly non-relevant terms (IT). Notice that non-relevant terms (iT) correspond to stop words. Each text is finally represented by these sets of terms, weights being associated with each term.

2.1.1 Text processing

Texts are processed using the following method:

1. Stop words are removed,
2. The remaining words are normalized using a dictionary that provides a common root for different words. This dictionary contains 21291 entries.
3. Alternatively phrases are extracted. Phrases correspond to frequent sequences of words or frequent sequences of word roots.

2.1.2 Topic processing

A topic is pre-processed in order to mark-up the sentences that describe the information relevancy and the sentences that describe the non-relevancy (see Figure 1: NarrativeRel and NarrativeNonRel tags).

Topic: 35
Title: NATO, Poland, Czech Republic, Hungary
Type: event
Descriptive: Accession of new NATO members: Poland, Czech Republic, Hungary, in
1999.
NarrativeRel: Identity of current and newly-invited members, statements of support for and opposition to NATO enlargement and steps in the accession process and related special events are relevant. Impact on the new members, i.e., requirements they must satisfy, and their expectations regarding the implications for them are relevant. Progress in the ratification process is relevant.
NarrativeNonRel: Future plans for NATO expansion, identification of nations admitted on previous occasions, and comments on future NATO structure or strategy are not relevant.

Figure 1: topic 35 (TREC 2003)

Then it is analyzed in order to extract the representative terms (words or phrases) as explained in the previous section. Each term is then weighted and categorized into 3 groups:

- Highly relevant terms are terms that get a weight greater than τ_H ,
- Scarcely relevant terms are terms that get a weight equal to τ_L ,
- Highly non-relevant terms are terms that are associated with non-relevancy in the narrative part of the documents.

More precisely, the formula used to compute the term weights is defined as follows:

Given Q_k a topic and t_i a term, $T_k = \{t_i \in Q_k / t_i \textit{ is not a stop word}\}$

T_k = TT_k U TD_k U TNR_k U TNN_k where TT_k corresponds to the set of terms extracted from the Title of the topic, TD_k from the Descriptive, TNR_k from the NarrativeRel and TNN_k is the NarrativeNonRel topic part.

$tf_{i,k,P}$ is the frequency of t_i in the TP_k part, $P \in \{T, D, NP, NN\}$

The term weight regarding a topic is computed as follows:

$$\begin{aligned}
&\omega_{1,i,k} = \sum_{P\in\{T,D,NP\}} \mu_P \cdot tf_{i,k,P} \\
&\omega_{2,i,k} = \mu_{NN} \cdot tf_{i,k,NN} \\
&\omega_{i,k} = \omega_{1,i,k} + f(\omega_1, \mu_{NN}) \cdot \omega_{2,i,k} \qquad where \quad f(\omega_1, \mu_{NN}) = 0 \quad if \quad \omega_{1,i,k} > 0 \; and \; \mu_{NN} < 0 \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad\;\; = 1 \qquad otherwise
\end{aligned}$$

$$\begin{aligned}
weight(t_i, Q_k) &= \omega_{i,k} \quad if \quad \omega_{i,k} \geq \tau_H \\
&= \omega_{2,i,k} \quad if \quad \omega_{1,i,k} = 0 \\
&= \tau_L \quad if \quad 0 < \omega_{i,k} < \tau_H \\
&= 0 \qquad otherwise
\end{aligned}$$

τ_L and τ_H are used in order to obtain a significant difference -in terms of importance- between highly relevant terms and scarcely relevant terms. The weights associated to scarcely relevant terms are set to τ_L (1 in the experiments submitted to TREC). τ_H is set to 3 in the TREC runs. This formula is also used in order to take into account highly non-relevant terms.

The term weight is used to categorized a term into one of the groups defined as follows:

$$\begin{gathered}
HT_k = \{t_i / t_i \in \bigcup\{TT_k, TD_k, TNR_k\} \; and \; weight(t_i, T_k) > \tau_L\} \\
LT_k = \{t_i / t_i \in (TNN_k - \bigcup\{TT_k, TD_k, TNR_k\}) \; and \; weight(t_i, T_k) = \tau_L\} \\
iT_k = \{t_i / \; weight(t_i, TP_k) = 0 \quad \forall P \in \{T, D, NR, NN\}\} \\
IT_k = \{t_i / t_i \in TNN_k \; and \; weight(t_i, T_k) < 0\}
\end{gathered}$$

2.1.3 Document processing

Each sentence of a document is considered as a text and the representative terms are extracted as explained in the section 2.1.1. To each term is associated a weight defined as follows:

Given S_j a sentence, t_i a term and $tf_{i,j}$ is the frequency of t_i in S_j. $\qquad weight(t_i, S_j) = tf_{i,j}$

2.2 Relevant sentences

In order to decide if a sentence is relevant, we associate three components to each sentence:

- a score that reflect the sentence – topic matching :

Given a topic Q_k and a sentence S_j

$$Score\,(S_j, Q_k) = \sum \left(weight(t_i, S_j) \cdot weight(t_i, Q_k) \right)$$

- and two groups of terms:

$$\begin{gathered}
HS_j = \{t_i / t_i \in (Sj \cap HT_k)\} \\
LS_j = \{t_i / t_i \in (Sj \cap LT_k)\}
\end{gathered}$$

HS_j corresponds to the highly relevant terms from the topic that also occurs in the sentence,

LS_j corresponds to the scarcely relevant terms from the topic that also occurs in the sentence.

Note that IT_k and iT_k sets are only used to calculate the term weight and it is not used in the sentence selection process.

A given sentence S_j is then considered as relevant iff :

$$Score\,(S_j, Q_k) > f\left(\frac{|LS_j|}{|LS_j| + |HS_j|}\right)\cdot|HT_k| + g\left(\frac{|HS_j|}{|LS_j| + |HS_j|}\right)\cdot|LT_k|$$

where $|X|$ is the number of elements of X

In the experiments that correspond to the runs sent to TREC, the function $f(\,)$ and $g(\,)$ have been set to:

$$f(x) = 2 - 1.5\,x \qquad and \qquad g(x) = 0.85 - 0.5\,x$$

2.3 Novel sentences

To decide if a sentence p is to be considered as novel, we compute the similarity between the sentence p and the previous successfully processed sentences p_i (novel) and the similarity between the sentence p and a sentence P' automatically built from the union of the set of p_i:

Given

- $\Pi = \{p_1, p_2, \ldots, p_n\}$ a set of sentences labeled as novel and $P' = \bigcup_{i\in\{1,\ldots,n\}} p_i$, P' is a sentence made of the set of sentences from Π,
- $Sim(x, y\,)$ a function that compute a similarity between x and y and
- p a sentence for which the system has to decide if it brings new information.

We first compute the following similarities:

$Sim(p, P') = \alpha_p$ and for $i \in \{1,\ldots,n\}$ $Sim(p, p_i) = \omega_{p,i}$

We then consider the q best previous sentences:

$for\ i \in \{1,\ldots,n\}$ $P_{p,i}$ is the series of sentences obtained by ordering Π in decreasing order of $\omega_{p,i}$.

$\beta_p = \sum_{i\in\{1,\ldots,q\}} Sim(p, P_{p,i})$ where $q \in \{4,5\}$ in the runs sent to TREC.

p is considered as redundant (not novel) iff:

$$\alpha_p \geq \tau_1 \text{ and } \beta_p \geq \tau_2$$

where $\tau_1 = 1$ and $\tau_2 = 0.6$ for the runs sent to TREC.

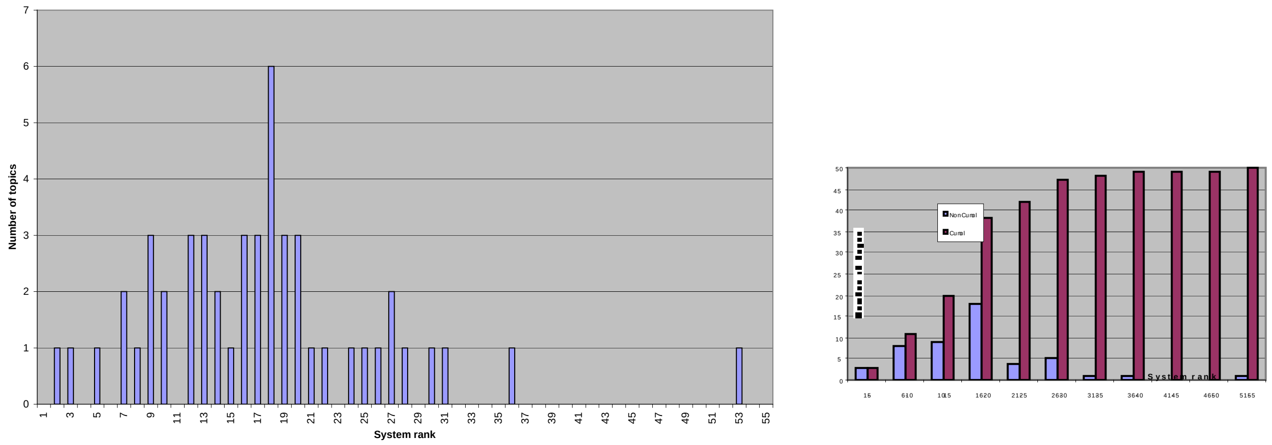
3 Results

This section presents the results we obtained with the method we developed and using the parameters as described in section 2. When comparing the results with the other runs, we can notice that our system is better in finding relevant sentences than in detecting novelty in the sentences. The difficulties of our system to detect novelty can be linked to the fact that the system does not take into account the order of the sentences in the documents.

3.1 Relevant sentences

Figure 1 indicates the number of topics for which our best system (or run) has been ranked at the Xth position among the 55 runs according to the F-Measure. For example, our method obtains the best results for 0 topic, the second position for 1 topics, the third for 1 topics, etc. and has a rank higher than 36th for only one topic (see figure 1.a). Figure 1.b provides a graph that summarizes figure 1.a by grouping together the results obtained for ranges of ranks. Additionally, the cumulative number of topics per range of system position is provided on the same graph. For example, we obtained a rank between 1 to 5 for 3 topics. The system obtains a rank equal or higher than 20 for 38 topics.

This clearly shows that our method is better than the average of the results. To be more precise, over the 50 topics, we obtained 46 topics (92%) for which the F-measure is higher or equal to the average of the 55 runs. And if we consider the run ranks, we obtained a rank higher or equal to the median (27) for 42 topics (84%).



a) Number of topics per run rank : detailed results

b) Number of topics per run rank : summarized results

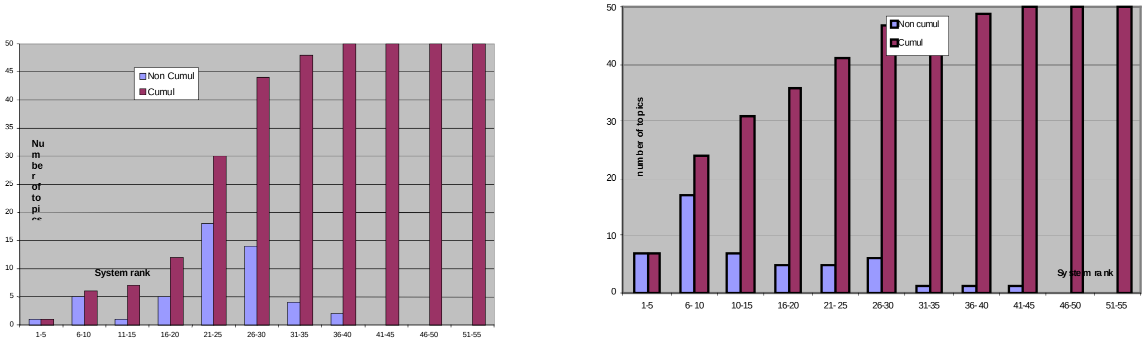
Figure 1: Number of topics per run rank – relevant sentences

3.2 New sentences

We present the results obtained in the second subtask the same way (see Figure 2). We distinguish the results when novel sentences are extracted from the retrieved sentences (TREC task 1 ; figure 2.1) and when they are extracted from the set of known relevant sentences (TREC task 2, figure 2.1).

Regarding the first case, over the 50 topics, we obtained 46 topics (92%) for which the F-measure is higher or equal to the average of the 55 runs. And if we consider the run ranks, we obtained a rank higher than the median (27) for 41 topics (82%).

However, when considering the relevant sentences, over the 50 topics, we obtained 41 topics (82%) for which the F-measure is higher or equal to the average of the 55 runs. And if we consider the run ranks, we obtained a rank higher than the median (27) for 30 topics (60%).



a) Novelty from retrieved sentences

b) Novelty from relevant sentences

Figure 2: Number of topics per run rank – summarized results

3.3 Other results

We modified the term weighting function in order to take better into account the query part in which the term occurs. The best results for the *relevant* subtask we obtained are the following: over the 50 topics, we obtained 46 topics (92%) for which the F-measure is higher or equal to the average of the 55 runs. And if we consider the run ranks, we obtained a rank higher or equal to the median (27) for 45 topics (90%). For one topic we obtained the best rank.

4 Conclusion

The approach we developed leads to relevant results for the first part of the task (relevant sentences). Over the 50 topics, we obtained 46 topics (92%) for which the F-Measure is higher or equal to the average of the 55 runs. And if we consider the run ranks, we obtained a rank higher than the middle (26) for 42 topics. Our best-submitted run obtains the following results: Average precision 0.64, Average recall 0.58 and Average F 0.526. With regard to the second sub-task (novelty), the submitted results over the 50 topics, we obtained 41 topics (82%) for which the F-measure is higher or equal to the average of the 55 runs.

5 References

[trec.nist.gov] TREC web site.

(Dkaki et al., 2002) T. Dkaki, J. Mothe, J. Augé, Novelty track at IRIT-SIG, Text Retrieval Conference TREC 2002, pp 332-336, 2003.

(Luhn, 1960) Luhn, H., Keyword in Context Index for Technical Literature, American Documentation XI (4), 1960, 288-295.

(Mothe et al., 2002) J. Mothe., C. Chrisment, B. Dousset, J. Alaux, DocCube : multi-dimensional visualisation and exploration of large document sets, Journal of the American Society for Information Science and Technology, JASIST, Special topic section: web retrieval and mining, Guest Editor: Hsinchun Chen, 54 (7), pp. 650-659, March 2003.

(Corral, 1995) M-L. Corral, J. Mothe, How to retrieve and display long structured documents*, Basque International Worshop on Information Technology*, pp 10-19, 1995