NOVELTY TRACK AT IRIT – SIG

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

(1) Institut de Recherche en Informatique de Toulouse, 118 Rte de Narbonne, 31062 Toulouse CEDEX

(2) IUT URS, 72 Rte du Rhin, 67400 Strasbourg

(3) Institut Universitaire de Formation des Maîtres Midi-Pyrénées, 56 av de l'URSS, 31078 Toulouse CEDEX

Abstract

IRIT developed a new strategy in order to detect the *relevant sentences* that we did not try in a more general context of document retrieval but did try previously and partially in document categorization. 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 have been defined: highly relevant, lowly relevant and no relevant. With regard to the *novelty part*, a sentence is considered as novel when its levels of coverage with the previously processed sentences and with the best-matching sentences do not exceed certain thresholds.

1 Introduction

«The TREC 2002 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 representation 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. IRIT developed a new strategy in order to detect the relevant sentence that we did not try in a more general context of document retrieval but did try previously and partially in document categorization. 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 have been defined: highly relevant, lowly relevant and no relevant. With regard to the novelty part, a sentence is considered as novel when its levels of coverage with the previously processed sentences and with the best-matching sentences do not exceed certain thresholds.

The results we obtain are quite good for the 'relevant' part. Indeed, we obtain 36 topics (73%) for which the R*P is higher or equal to the average of the 42 runs. With regard to the 'novelty' part, the results are disappointing and our method is situated around the average.

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). In section 3 we present the results and comment them. Finally we indicate in the last section the future directions for our work on novelty track.

2 Description of the method

2.1 Document and topic representation

In our method, topics and sentences are considered as texts. Each text is pre-processed the same way in order to extract the representative terms. Then, the terms extracted from the topics are categorized into two groups : highly relevant terms (HT) and lowly relevant terms (LT). Finally, each text is represented by these two set of terms, with weights associated to 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.

2.1.2 Topic processing

A topic is considered as a single text and the representative terms are extracted as explained in the previous section. Each term is then weighted and categorized into 2 groups:

- Highly relevant terms are terms that get a weight greater than 3,
- Lowly relevant terms are terms that get a weight equal to 1 (see below the formula used to compute the term weights).

Given T_k a topic, t_i a term and $tf_{i,k}$ the frequency of t_i in T_k.

The term weight regarding a topic is computed as follows:

$$\begin{aligned} weight(t_i, T_k) &= tf_{i,k} \qquad if \quad tf_{i,k} \geq 3 \\ &= 1 \qquad otherwise \end{aligned}$$

In order to obtain a significant difference -in terms of importance- between the highly relevant terms and the lowly relevant terms the weights of the lowly relevant terms are set to 1.

Each term is also categorized into one of the groups defined as follows:

$$HT_k = \{t_i / t_i \in T_k \ and \ weight(t_i, T_k) > 1\}$$

$$LT_k = \{t_i / t_i \in T_k \ and \ weight(t_i, T_k) = 1\}$$

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.

$$weight(t_i, S_j) = tf_{i,j}$$

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 T_k and a sentence S_j

$$\begin{aligned} Score(S_j, T_k) &= \sum \left(weight(t_i, S_j) \cdot weight(t_i, T_k) \right) \\ &= \sum_{t_i / t_i \in HT_k} \left(tf_{i,j} \cdot tf_{i,k} \right) + \sum_{t_i / t_i \in LT_k} \left(tf_{i,j} \right) \end{aligned}$$

- and two groups of terms:

$$HS_j = \{t_i / t_i \in (Sj \cap HT_k)\}$$

$$LS_j = \{t_i / t_i \in (Sj \cap LT_k)\}$$

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

LS_j corresponds to the lowly relevant terms from the topic that occurs in the sentence,

A given sentence S_j is then considered as relevant iff :

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

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

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

$$f(0) = 2 \; ; \forall x \in]0,1], \; f(x) = 1.5$$
$$g(0) = 0.85 \; ; \forall x \in]0,1], \; g(x) = 0.3$$

2.3 Novelty sentences

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

Given

$P = \{p_1, p_2, \ldots, p_n\}$ a set of sentences and $P' = \bigcup_{i \in \{1,\ldots,n\}} p_i$, P' is a sentence made of the set of sentences P ,

$Sim(x, y)$ a function that compute a similarity between x and y and

p a sentence for which the system has to decide if it brings something new.

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:

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

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

p is considered as novel iff:

$$\alpha_p \geq T_1 \text{ and } \beta_p \geq T_2$$

where $T_1 = 1$ and $T_2 = 0.6$ for the run sent to TREC.

3 Results

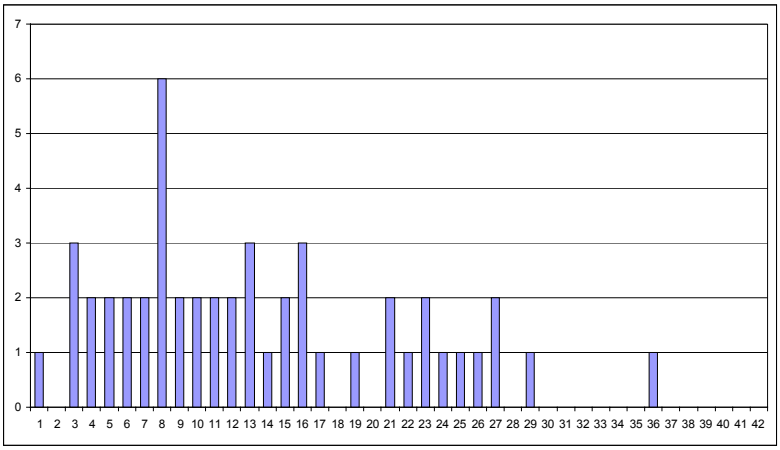
This section presents the results we obtained with the method we developed and using the parameters that have been described in section 2.

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. This can be explained by the method we used that does not take into account the order of the sentences in the documents. Additionally most of the parameters have to be tune specifically to take into account the sentence relationship.

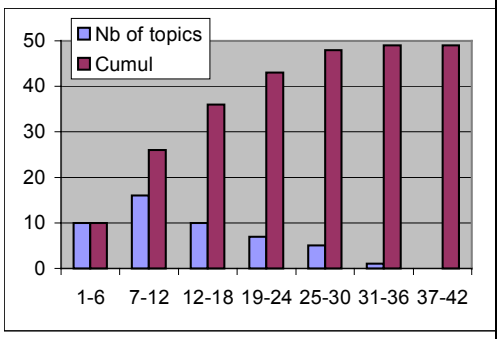
3.1 Relevant sentences

Figure 1 indicates the number of topics for which our system (or run) has been ranked at the X^{th} position among the 42 runs. For example, our method obtains the best results for 1 topic, the third position for 3 topics, the fourth for 2 topics, etc. and has a rank higher than 30^{th} for only one topic (see figure 1.a). Figure 1.b provides a graph that summarize 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 6 for 10 topics. The system obtains a rank equal or higher than 24 for 43 topics.

This clearly shows that our method is better than the average of the results. To be more precise, over the 49 topics, we obtained 36 topics (73%) for which the R*P is higher or equal to the average of the 42 runs. And if we consider the run ranks, we obtained a rank higher than the middle (21) for between 37 and 39 topics (depending how we consider the rank when 2 systems obtained the same value for R*P).



a) Number of topics per run rank : detailed results



b) Number of topics per run rank : summarized results

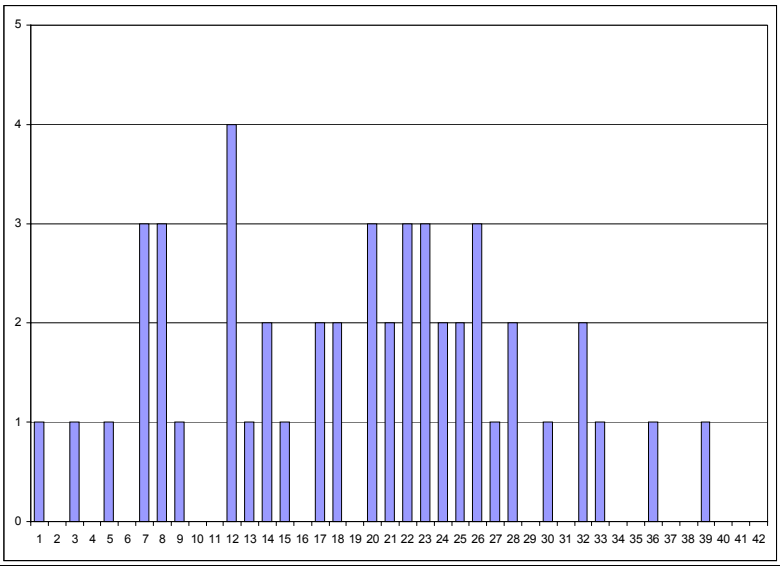
Figure 1 : Number of topics per run rank – relevant sentences

3.2 New sentences

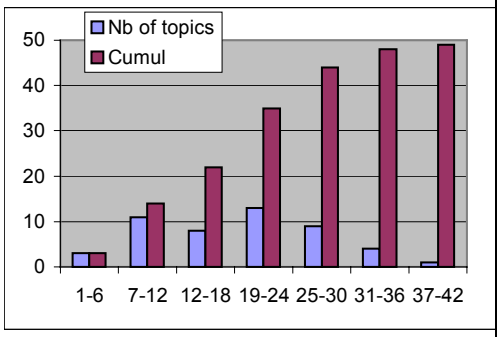
We present the results obtained in the second subtask the same way (see Figure 2).

Over the 49 topics, we obtained 20 topics (about 40%) for which the R*P is higher or equal to the average of the 42 runs. And if we consider the run ranks, we obtained a rank higher than the middle (21) for between 23 and 27 topics (depending how we consider the rank when 2 systems obtained the same value for R*P). The method is not better than the average.

The results in the second subtask are directly linked to the results obtained in the first subtask. As a result, the groups that obtained high R*P in the first part are more likely to obtain good results in the 'novelty' part. We can consider that the results we obtained for the novelty part are quite disappointing as the results on the relevance part were good. One of the tracks we are going to explore to improve the results is to take into account the order of the sentences as two sentences from a same paragraph are more likely to treat the same subject for example.



a) Number of topics per run rank : detailed results



b) Number of topics per run rank : summarized results

Figure 2 : Number of topics per run rank - novelty

4 Conclusion

The approach we developed leads to relevant results for the first part of the task (relevant sentences). Over the 49 topics, we obtained 36 topics (73%) for which the R*P is higher or equal to the average of the 42 runs. And if we consider the run ranks, we obtained a rank higher than the middle (21) for between 37 and 39 topics. With regard to this sub-task, future work will be devoted first improving the definition of the $f(\,)$ and $g(\,)$ functions, which play the role of thresholds.

With regard to the second sub-task (novelty), the results are just on the average. This can be explained by the method we used that does not take into account the order of the sentences in the documents. Additionally most of the parameters have to be tune specifically to take into account the sentence relationship (two sentences that are close together in a document are more likely to deal with the same subject). This probably will also improve the fist sub-task of novelty track.

5 References

[trec.nist.gov] TREC web site.

[Luhn, 60] Luhn, H., Keyword in Context Index for Technical Literature, American Documentation XI (4), 1960, 288-295.

[Mothe, 02] Mothe J., Chrisment C., Dousset B., Alaux J., DocCube : multi-dimensional visualisation and exploration of large document sets, to appear in JASIST.