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 2004, IRIT modified important features of the strategy that was developed for TREC 2003. Changes include tuning parameter values, topic expansion and exploitation of sentences context.

According to our method, a sentence is considered as *relevant* if it matches the topic with a certain level of coverage. This coverage depends on the category of the terms used in the texts. Four types of terms have been defined highly relevant, scarcely relevant, non-relevant (like stop words), highly non-relevant terms (negative terms). Term categorization is based on topic analysis: highly non-relevant terms are extracted from the narrative parts that describe what will be a non-relevant document. The three other types of terms are extracted from the rest of the query. Each term of a topic is weighted according to both its occurrence and the topic part it belongs to (title, descriptive, narrative). Additionally we increase the score of a sentence when either the previous or the next sentence is relevant. When topic expansion is applied, terms from relevant sentences (task 3) or from the first retrieved sentences (task 1) are added to the initial terms.

With regard to the *novelty part*, a sentence is considered as novel if its similarity with each of 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 and previously selected sentences.

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 the general context of document retrieval but we did use it previously and partially in document categorization (Mothe, 2002) and XML retrieval (Hubert, 2005). 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 topics that describe what non-relevant documents are. A negative weight can be assigned to these words. In TREC 2004, IRIT modified important parameter features of this approach. This includes parameters retuning, topic expansion and exploitation of sentences context. When topic expansion is applied, terms from relevant

sentences (task 3) or from the first retrieved sentences (task 1) are added to the initial terms. The context of a sentence is taken into account by increasing the score of a sentence when either the previous or the next sentence is relevant.

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 and previously selected sentences. The similarity function is based on the dot-product function, vectors representing the sentences does not take into account neither the weight of the stems (we use Boolean vectors) nor the context in which the sentence occurs.

The rest of the 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 three tasks. In section 3 we present the results and give some comments. Finally, in section 4, we discuss our results and the evolution in Novelty Track results (2002 to 2004).

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. Terms extracted from a given topic are then 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. Extracted terms are weighted (see below). Each text is finally represented by these sets of weighted terms.

Note that the values of the different parameters in the formulas are given in section 3.

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 its sentences that describe relevant documents and the sentences that describe non-relevant documents (see Figure 1: NarrativeRel and NarrativeNonRel tags).

Topic: 59
Title: Payne Steward Plane Crash
Type: Event
Descriptive: Identify a document that describes the plane crash that killed the golfer Payne Stewart on Oct. 25, 1999.
Narrative: Details about the crash, who else was aboard, and information about the destination and departure are relevant. The reason for the flight would not be relevant. Time and weather conditions are relevant.

NarrativeRel: Details about the crash, who else was aboard, and information about the destination and departure are relevant. Time and weather conditions are relevant.
NarrativeNonRel: The reason for the flight would not be relevant.

Figure 1: topic 59 (TREC 2004)

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 one of the 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, $T_k = \{t_i \in Q_k / t_i \text{ is not a stop word}\}$

$T_k = TT_k \cup TD_k \cup TNR_k \cup TNN_k$ 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:

$$\omega_{1,i,k} = \sum_{P \in \{T,D,NP\}} \mu_P \cdot tf_{i,k,P}$$

$$\omega_{2,i,k} = \mu_{NN} \cdot tf_{i,k,NN}$$

$$\omega_{i,k} = \omega_{1,i,k} + f(\omega_1, \mu_{NN}) \cdot \omega_{2,i,k} \qquad where \quad f(\omega_1, \mu_{NN}) = 0 \quad if \quad \omega_{1,i,k} > 0 \text{ and } \mu_{NN} < 0$$
$$= 1 \qquad otherwise$$

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

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. 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 categorize a term into one of the following groups:

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

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

2.2.1 Without topic expansion

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:

$$HS_j = \{ P \in \{T, D, NR, NN\} \}$$
$$LS_j = \{ t_i / t_i \in (Sj \cap LT_k) \}$$

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.

A given sentence S_j is then considered as relevant iff :

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

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

2.2.2 With topic expansion

Topic expansion is either based on blind relevance feedback using the first retrieved sentences (task 1) or relevance feedback using the sentences known as relevant (task 3). In both cases the model takes into account topic expansion using the following formula:

$$\omega_{i,k} = \sum_{P \in \{T, D, NP, NN1, NN2\}} \mu_P \cdot tf_{i,k,P} + \mu_{RP} \quad (5)$$

when $tf_{i,k,P} > \Delta$

Where RP is the set of sentences used for topic expansion and Δ is a constant used as a threshold.

The detection of relevant sentences is then based on the formulas described section 2.2.1.

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 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 all the sentences from Π,
- $Sim(x, y)$ a function that computes the 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 $Sim(p, p_i) = \omega_{p,i}$ for $i \in \{1, \ldots, n\}$

We then consider the q best matching sentences:

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

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

p is considered as redundant (not novel) iff:
$\alpha_p \geq \tau_1$ and $\beta_p \geq \tau_2$
$\tau_1 = 1$ 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 as described in section 2.

3.1 Description of the runs IRIT submitted

	Name	Description
TASK 1	IRITT1	411-1622-10.NormBonif1.10.4
	IRITT2	411062200.NormBonif1.15.2
	IRITT3	411062200.FuncRetroBonif1.10.4
	IRITT4	411062200.FuncRetroBonif1
	IRITT5	411062200.NormBonif1
TASK 2	**IritTask2**	Uses IRITT3 and the process explained 2.3
	Irit2T2	All relevant sentences are considered as novel
TASK 3	Irit1T3	411062200.Func2.15.4
	Irit2Task3	411062200.Func2Bonif1.15.4
	Irit3Task3	411062200.FuncBonif1.10.4
	Irit4Task3	411062200.FuncRetroBonif1.15.4
	Irit5Task3	411-1622-10.Func2Bonif1.15.4

Figure 2: Description of the runs IRIT submitted. The best run for each task is in bold characters.

The description provides the value of the different parameters of our method:

- The first series of values corresponds to the coefficients associated to the different topic parts and that are used to define the class of each extracted term (see section 2.1.2). μT, μD, μNR, μNN, first for terms then for phrases.
- The second part of the description indicates the function we used to select relevant sentences

Func corresponds to $f(x) = 2 - 1.5\,x$ *and* $g(x) = 0.85 - 0.5\,x$, $\alpha = 0$

Norm corresponds to $f(x)=g(x)=0$, $\alpha=3$

Bonif1: means that we increase by 1 the score of a sentence that follows a relevant sentence

BonifRetro: means that we increase by 2 a sentence that follows a relevant sentence and by 1 a sentence that precede a relevant sentence.

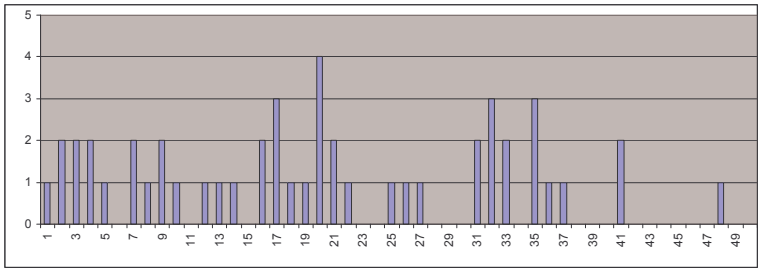
- The third and fourth parts correspond to the maximum number of sentences that are taken into account for feedback and the value of Δ (see section 2.2.2).

3.2 Relevant sentences

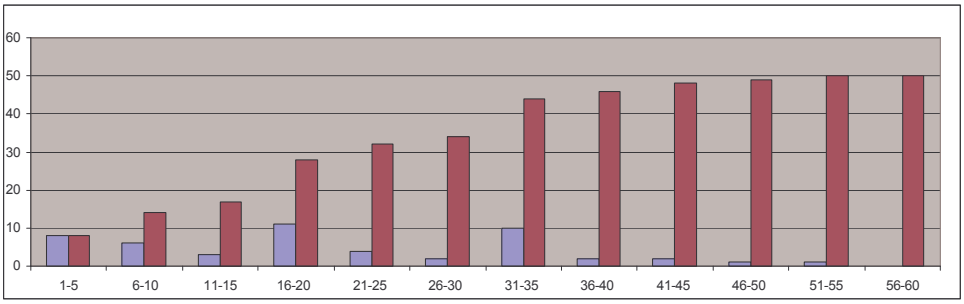
3.2.1 Task1

Figure 3 indicates the number of topics for which our best system (or run) has been ranked at the X^{th} position among the 60 runs according to F-Measure. For example, our method obtains the best results for 1 topic, the second position for 2 topics, the third for 2 topics, etc. and has a rank higher than 41^{th} for only two topic (see figure 3.a). Figure 3.b provides a graph that summarizes figure 3.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 and 5 for 8 topics. The system obtains a rank equal or lower to 20 for 28 topics.

This clearly shows that our method is better than average. To be more precise, over the 50 topics, we obtained 41 topics (82%) for which F-measure is higher or equal to the average F-measure over the 60 runs. And if we consider the run ranks, we obtained a rank higher or equal to the median (30) for 34 topics. There is no correlation between these results and the type of topic (event or opinion).



a) Number of topics per run rank : detailed results



b) Number of topics per run rank : summarized results

Figure 3: Number of topics per run rank – relevant sentences

3.2.2 Task 3

Our training method seems to be insufficiently efficient compared to others' as our system ranked at a lower position when trained. On average, our system ranked at the 20th position (over 60 runs) without training and at the 23rd position (over 40 runs) when trained. Our system obtained a F-measure higher to the average for 29 topics (against 41 without training). However the results are different for event and opinion topics. Among the 29 topics, 12 are events and 17 are opinions.

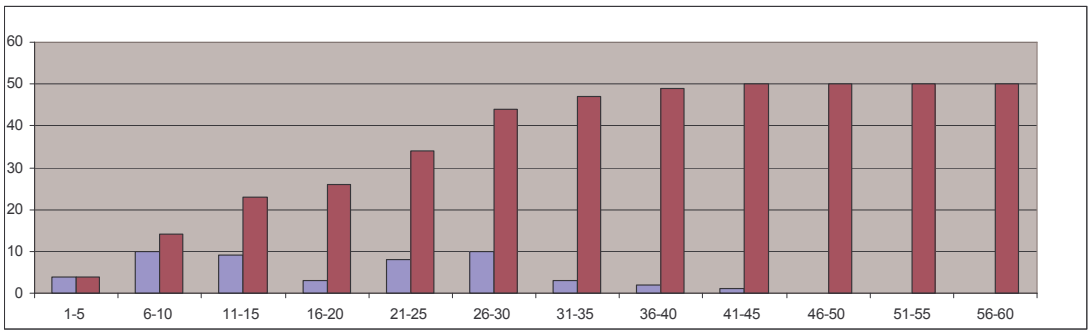
3.3 New sentences

We present the results obtained when detecting novel sentences the same way (see Figure 4). We distinguish the results when novel sentences are extracted from the retrieved sentences (TREC task 1 ; figure 4.a) and when they are extracted from the set of sentences known as relevant (TREC task 2, figure 4.b).

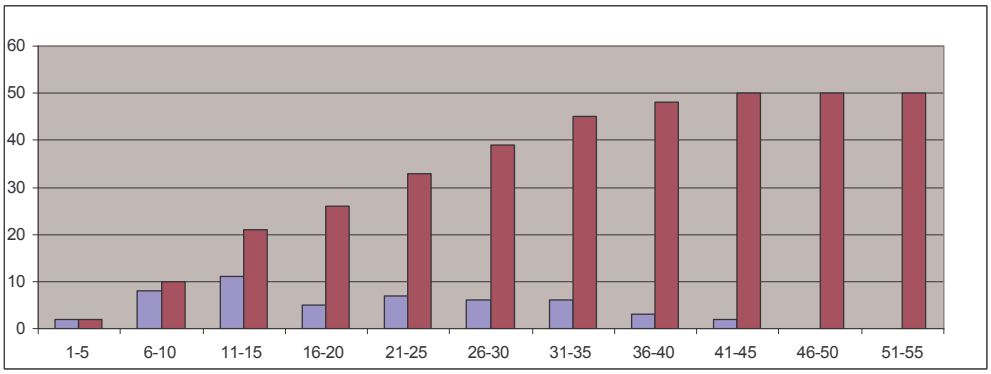
Regarding the first case (task 1), over the 50 topics, we obtained 43 topics (86%) for which F-measure is higher or equal to the average over the 60 runs. And if we consider the run ranks, we obtained a rank higher than the median (30) for 44 topics (88%).

However, when considering the relevant sentences (task 2), over the 50 topics, we obtained 46 topics (92%) for which F-measure is higher or equal to the average over the 55 runs. Event and opinion topics are evenly distributed (23-23). If we consider the run ranks, we obtained a rank higher than the median (27) for 34 topics (68%). In that case, events and opinions are not evenly distributed (22-12).

The results obtained in Task 3 at the *new* sub-task, are quite comparable to those obtained at the *relevance* sub-task (i.e. our system under-performed comparatively to the other systems). This probably does not question the method we define to detect new sentences since it was carried out on a 'noisy' set of sentences (We do not detect relevant sentences well, as a result, F-measure is low for novelty detection.)



a) Novelty from retrieved sentences (task 1)



b) Novelty from relevant sentences (task 2)

Figure 4: Number of topics per run rank – summarized results

4 Discussions

The results we obtained in TREC 2002 were quite good regarding the 'relevant' subtask. Indeed, for 36 topics (73%), R*P measure was higher or equal to F-measure averaged over the 42 runs that were submitted.

In TREC 2003, we improved these results as we obtained 46 topics (92%) for which F-measure (2*R*P/(R+P)) was equal or higher to F-measure averaged over the 55 runs submitted. With regard to the 'novelty' part, we also obtained 46 topics (92%) for which F-measure was higher or equal to the average over the 55 runs. An interesting fact is that, relatively to other participants' methods, our method performed better when there was some 'noise' in the sentence set. Indeed the results were better when considering the retrieved sentences than when considering only the relevant sentences, (i.e. our system ranked better over the submitted runs in the case of 'noisy' sentences). We obtained 41 topics (82%) for which F-measure was higher or equal to the average over the 55 runs.

In TREC 2004, regarding task1 and the detection of the sentence relevancy, for 41 topics (82%), F-measure is higher or equal to F-measure averaged over the 60 submitted runs. 20 of these 41 topics are "events" and 21 are "opinions" (thus event and opinion topics are evenly distributed among these 41 topics).

With regard to the 'novelty' part (task 1), when considering the retrieved sentences, 43 topics get F-measure higher or equal to F-measure averaged over the runs. However, in this case, the distribution is slightly different, as there are 23 events and 20 opinion topics. Regarding task 2,

for which only relevant sentences are considered to detect novelty, 46 topics (92%) get F-measure equal or higher than F-measure averaged over the 55 runs (evenly distributed among event and opinion topics). A system that would consider all relevant sentences as novel would get only 29 topics for which F-measure would be higher or equal to average (evenly distributed among event and opinion topics). This is far less than the average performance of all tested systems. This was not the case in 2002 and 2003.

When some information (relevant sentences) is used for learning purpose (task 3), our system detects relevant sentences better than the average over the 40 systems (runs) for 29 topics and it detects novel sentences better than the average for only 6 topics. However, up to 19 topics are better than the median for the latter sub-task. This means that runs either performed very well or very badly (big standard deviation).

Last year the system was clearly better detecting relevance than novelty. In TREC 2004, this difference is no clear any more. The results we obtained are better than results averaged over the runs. Regarding relevance detection (task 1), our best run obtains the following results: Average precision 0.32, Average recall 0.74 and Average F-measure 0.404. With regard to the novelty detection (task 1), our best run obtains the following results: Average precision 0.15, Average recall 0.68 and Average F-measure 0.221. When using relevant sentences only, the novelty detection process obtains: Average precision 0.45, Average recall 0.98 and Average F-measure 0.605. Finally, when using relevance judgments to *train* the system to detect relevance, the system obtains: Average precision 0.29, Average recall 0.68 and Average F-measure 0.372.

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.

(Dkaki et al. 2004) T. Dkaki, J. Mothe, Combining Positive and Negative Query Feedback in Passage Retrieval, RIAO, pp 661-672, 2004. http://www.irit.fr/~Josiane.Mothe

(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.