Document Re-ranking Based on Topic-Comment Structure

Liana Ermakova, Josiane Mothe

Document Re-ranking Based on Topic-Comment Structure

Liana Ermakova
Institut de Recherche en Informatique de Toulouse
France
Email: ermakova@irit.fr

Josiane Mothe
Institut de Recherche en Informatique de Toulouse
France
Email: mothe@irit.fr

***Abstract*—This paper introduces a novel approach for document re-ranking in information retrieval based on topic-comment structure of texts. While most information retrieval models make the assumption that relevant documents are about the query and that aboutness can be captured considering bags of words only, we rather consider a more sophisticated analysis of discourse to capture document relevance by distinguishing the topic of a text from what is said about the topic (comment) in the text. The topic-comment structure of texts is extracted automatically from the first retrieved documents which are then re-ranked so that the top documents are the ones that share their topics with the query. The evaluation on TREC collections shows that the method significantly improves the retrieval performance.**

***Index Terms*—Information retrieval, document re-ranking, information structure, topic, comment, theme, rheme.**

I. Introduction

Information retrieval (IR) is usually grounded on the hypothesis that relevant documents are *about* the query; the query being supposed to reflect properly the user's information need [1].

Aboutness is not as simple to define as it seems and IR suggested various definitions. For example, Cummins [2] mentions that the term-occurrence frequency is "a measure of the degree to which a document is about a specific term". Concretely, most of IR models make the hypothesis that aboutness can be caught by matching the query terms and the document terms, both considered as bags of words [3][1]. Aboutness is thus seen at a general level, considering the discourse topic, that is to say what the entire text or paragraph (in case of focused or XML passage retrieval) is about.

In linguistics, the notion of aboutness is more complex and is related to the **topic** (or **theme**), which is what the text (typically a sentence) is about, while the **comment** (or **rheme** or **focus**) is what is being said about the topic [4].

As a matter of fact, when seeking for information using a search engine, the user is generally interested by the comment not by the topic. Although, the topic is mandatory to make the link between the user's information need and the text aboutness. Current IR models do not distinguish these two aspects in texts.

In this paper, our goal is to improve the ranking of retrieved document by taking advantage of the information structure, i.e. the topic-comment structure of texts. More precisely, in our approach the notion of aboutness is first considered at the discourse-level using current IR model and then at the clause level in order to re-order the retrieved documents so that the top ones are more likely to bring useful comments on the query topic. According to our model, rather than matching the query terms with the document terms wherever they occur in the information structure, we promote an approach in which the query terms should match differently the topic and the comment parts of the sentences.

Let consider a query *Dostoyevsky* and two examples of documents.

Example 1:
$\{\text{Dostoyevsky}\}_{topic}$ {expressed religious, psychological and philosophical ideas in his writings}$_{comment}$.
$\{\text{He}\}_{topic}$ {admired Hoffmann who influenced his works}$_{comment}$.

Example 2:
$\{\text{Berdyaev}\}_{topic}$ {expressed religious, psychological and philosophical ideas in his writings}$_{comment}$.
$\{\text{He}\}_{topic}$ {admired Dostoyevsky who influenced his works}$_{comment}$.

Example 1 is talking about Dostoyevsky's work while the second document (example 2) is about Berdyaev.

The traditional bag-of-words approaches are not able to distinguish the difference between these texts. Both documents would have the same score according to bag-of-words based methods since

- the query term *Dostoyevsky* occurs once in each document;
- the documents are of the same length;
- the only different terms are *Hoffmann* and *Berdyaev*.

In contrast to this, we hypothesize that document topics should occur in topic parts of sentences.

In most languages the common means to mark topic-comment relations are word order and intonation. However, since we are considering only textual documents in this study, we do not look at intonation annotation. In texts, the prominent construction for topic-comment is the so-called *topic fronting*. Topic fronting refers to placing the topic at the beginning of a clause regardless whether it is marked or not [4][5]. Thus, even if complex linguistic-based methods could be used to extract topic-comment structure from sentences, the topic fronting feature can be used as a simpler way to extract the information structure. Moreover too sophisticated linguistic

methods would not be applicable at a large scale to analyze document sentences for IR purposes.

In this paper, we focus on automatic annotation based on the topic fronting assumption. The method we proposed requires only shallow parsing, namely sentence chunking and part-of-speech (POS) tagging to automatically extract the information structure. Topic-comment identification can be either done off-line on the all collection or on-line on the retrieved document set. In the first case our approach could be applied as a ranking method. Since we applied topic-comment detection on the retrieved document set only, we use it as a re-ranking method.

We evaluate our method on two different collections: TREC Robust and WT10G. We compare our method considering several commonly used measures (MAP, $NDCG$ and $BPREF$) both to $BM25$ and a strong baseline consisting of an initial retrieval performed by Divergence from Randomness model $InL2$ and the $Bo2$ pseudo-relevance feedback method implemented in Terrier platform which provides state-of-the-art effective retrieval mechanisms [6].

The rest of the paper is organized as follows. Section II describes related works considering both topic-comment structure research and its applications in IR. Section III provides the novel method we promote for document re-ranking that exploits the information structure to better match queries and documents. Section IV describes the evaluation framework. Section V presents the results and discusses them. Section VI concludes the paper.

II. Related Work

A. Topic-comment Structure in Linguistics

Apparently, Henri Weil could be the one who introduced the topic-comment opposition in 1844 [7]. He established the connection between topic-comment structure and word order. At that time the topic was called a psychological subject, while the comment was defined as psychological predicate.

Definition 1: A clause-level topic is the phrase in a clause that the rest of the clause is understood to be about, and the comment is what is being said about the topic.

According to W. Mathesius [8], the topic does not provide new information but it connects the sentence to the context. Thus, the *topic* and the *comment* are opposed in terms of the given/new information. This contraposition is called **information structure** (i.e. the *topic-comment structure*).

Let's consider two examples:

Example 3:

$\{\text{Anna}\}_{topic}$ $\{\text{married Sam 3 years ago}\}_{comment}$.

Example 4:

$\{\text{Sam}\}_{topic}$ $\{\text{married Anna 3 years ago}\}_{comment}$.

The sentence in Example 3 is about Anna, while the sentence in Example 4 is about Sam. Thus, the topic of ex. 3 is Anna, while the topic ex. 4 is Sam. The comment is the answer on the question *What's about the topic?*

Topic-comment influence has been studied on speech technology. Research work investigates intonational focus assignment or the relation between discourse structure and posture and gesture in order to design embodied conversational agents.

Information structure in texts presupposes the dichotomy of information units, namely topic and comment [9]. These information units are triggers for syntactic and semantic processes, namely word order (dislocation), prosody ((de) accentuation), and interpretation. Dislocation and accentuation mainly appear within sentence bounds, while discourse linking put a sentence into a discourse context and thus influence the interpretation.

The collaborative research cluster (SFB) 632 proposed guidelines for the annotation of information structure [10] as follows:

Definition 2: A Noun Phrase (NP) X is the Aboutness Topic of a sentence S containing X if

1) S would be a natural continuation to the announcement **Let me tell you something about** X
2) S would be a good answer to the question **What about** X**?**
3) S could be naturally transformed into the sentence **Concerning** X**,** S^* where S^* differs from S only insofar as X has been replaced by a suitable pronoun.

Cook and Bildhauer [11] shows that despite using the same guideline, annotator agreement on topic-comment is sometimes difficult to obtain.

Actually, manual annotation of information structure in texts challenges the identification of the focus of a sentence or the discourse topic [12]. Versley and Gastel proposed to chunk texts into topic segments since the discourse relations are usually bounded by topic segments [12]. Relations (subordinating or coordinating) fall into the following categories: contingency, expansion, temporal, comparison, and reporting.

Some work has been carried out for automatic topic segmentation in broadcast news and has been applied for example in the Topic Detection and Tracking (TDT) program mainly based on word usage [13] or using prosodic clues [14].

Importantly enough, in texts, there exist special constructions to introduce the comment: topic fronting, placing the topic at the beginning of the clause is prominent. In this paper, rather than using discourse parser which is too time consuming for large amount of texts, we develop a simpler way of extracting topic-comment structure for IR (see Section III).

B. Discourse-level Topic vs Rhetorial Relations and Topic-comment Structure in IR

Matching the discourse-level topic referring to the notion of aboutness of a document has been well studied in IR literature [15][1][3]. However, modern search engines are essentially key word oriented and, thus, do not consider the relationships between terms [3] nor between topics [16]. On the other hand, linguistic analysis is crucial for text interpretation; as an example rhetorical relationships indicated how the parts of a coherent text are linked to each other.

Various parsers extract discourse structure such as HILDA [17] which implements topic changes or SPADE [18]. Both parsers were trained at the RST-DT corpus annotated according to Rhetorical Structure Theory [19]. Although the

original set of discourse relations were limited to 24, the RST-DT corpus contains about one hundred relations. This set is usually reduced by the integration of relations into classes. Thus, in SPADE discourse parser, 18 rhetorical relations are taken into account: attribution, background, cause-result, comparison, condition, consequence, contrast, elaboration, enablement, evaluation, explanation, manner-means, summary, temporal and topic-comment. However, the topic-comment relation in the RST-DT corpus (and therefore in SPADE and HILDA parsers) is defined in a different way. Indeed, we can find the following definition: topic-comment is "a general statement or topic of discussion is introduced, after which a specific remark is made on the statement or topic $\langle ... \rangle$ When the spans occur in the reverse order, with the comment preceding the topic, the relation comment-topic is selected. While comment-topic is not a frequently used mean in English, it is seen in news reporting, for example, when someone makes a statement, after which a reference is given to help the reader interpret the context of the statement $\langle ... \rangle$ Ex. [As far as the pound goes,] [some traders say a slide toward support at 1.5500 may be a favorable development for the dollar this week.]" [19]. These parsers are based on deep analysis of linguistic features and are hardly usable when large quantities of texts are involved. However, the major reason why we do not use a discourse parser to extract the topic-comment structure of texts is that the extracted topic-comment relation is not the same. Discourse parsers view the topic-comment relation as a remark on the statement while we consider a topic as the phrase that the rest of the clause is understood to be about.

Lioma et al. use rhetorical relations from SPADE parser to re-rank documents [20]. The authors introduced a query likelihood retrieval model based on the probability of generating the query terms from (1) a mixture of the probabilities of generating a query from a document and its rhetorical relations and (2) the probability of generating rhetorical relations from a document. One of the limitations of this approach is that not all types of texts can be parsed this way (e.g. legal texts or item lists have a few rhetorical relations). In addition, the rule-based parsers even if they take into account some statistics, are not extensible to other languages. An even more problematic drawback is related to the shortcomings of the discourse parser since such parsers are very time consuming and cannot be applied on large volumes of data. Lioma et al. state that topic-comment relations as defined by SPADE are extremely sparse in the benchmark IR collections [20], while in our approach topic-comment structure is common for all types of texts as well as for all genres.

Many other document re-ranking approaches consider user behavior, for example clicks or dwell time [21]. Some recent researches also take into account page view history [21]. Such approaches assumes multiple searches for the same information need. Li et al. introduced a document re-ranking using partial social tagging [22] which is the main limitation of the approach. Veningston and Shanmugalakshmi proposed to exploit term graph data structure and re-rank documents according to the association and similarity between them [23]. The authors stated that their approach involve expensive computation. Chou et al. suggested a Semantic Analysis on Relevance Feedback method for re-ranking which is a variant of topic modeling [24]. This approach may be considered as the bag-of-word based since it does not consider the relationships between words within a text.

In [25], the author proposed to exploit topic-comment structure for text summarization. There, the assumption of topic fronting was simplified by viewing a topic as the first half of a sentence. The author stated that topic-comment analysis did not improve results. A possible reason is the method of the topic-comment structure extraction. In contrast to [25], we propose to apply information structure for document re-ranking. Moreover, we introduce another algorithm for topic-comment chunking, namely we assume that a topic should be placed before a personal verb while the rest of the sentence is considered as a comment.

To the best of our knowledge, the closest related work is [26]. The authors propose to apply topic-comment structure for document classification while our approach aims at document re-ranking (but can be easily applied for document retrieval). They hypothesize that the important information belongs to the theme and that relevant documents to a query should share themes. The approach is underlain by the notions of topicality power and explanatory power that allows estimating document topicality by the cascade of neural networks. In contrast to this approach, we propose to integrate the topic-comment structure into the classical retrieval models such as $BM25F$ which is a variant of $BM25$ that takes into account document structure and multiple weighted fields. We choose $BM25F$ as a simplest and elegant way to assign different weights to different document parts. In contrast to $BM25F$ we do not use fields (structural components) but the set of the oppositions between topic and comment. Bouchachia and Mittermeir do consider only features within a document while we believe that it is important to take into account collection features. That is why we introduced the notion of Inversed Comment Frequency which is analogous of the concept of Inversed Document Frequency. The topic-comment annotation process in their approach requires syntax parsing, although other details are not provided in their paper.

III. Information Structure for Information Retrieval

A. Automatic Topic-comment Annotation

The topic-comment structure is opposed to formal structure with grammatical elements as the constituents. The difference between *topic* and grammatical subject is that topic refers to the information or pragmatic structure of a clause and how it is related to other clauses, while the subject is a merely grammatical category.

In simple English clause the topic usually coincides with the subject, even if it is not always the case as for expletives (e.g. *it is snowing*) that do not have topics at all [10]. Moreover, the unmarked word order in English is Subject - Verb - Object

(SVO). Thus, it is possible to make an assumption that, as a rule, the topic is placed before the verb. We make an additional assumption, that if a subordinate clause provides details on an object, it is rather related to the comment. Thus, the main idea of the proposed method is to split a sentence into two parts by a personal verb.

Here is an example of the topic-comment chunking from the TREC collection.

Example 5:
$\{$The Bengal Standard$\}_{topic}$ $\{$is a description of the ideal Bengal and therefore is used to define the quality of each cat$\}_{comment}$.

Our method requires only shallow parsing, namely sentence chunking and POS tagging. Even if this is a light NLP function, POS tagging can be a challenging issue if applied to an entire document collection. For that reason, we rather use the knowledge on information structure as a mean to re-rank documents that have been retrieved considering more traditional matching (e.g. BM25-based matching), although our algorithm is not limited to re-ranking.

The computational complexity of the proposed method for topic-comment identification is linear over the number of words.

B. Topic vs Comment for Query Matching

State-of-the-art models in IR consider the document ranking function as a matching function between the terms in the documents and the query without considering term relationships. In our model, we hypothesize that the topic-comment structure could be useful in the matching process. Moreover, we argue that topic matching would be more effective than term matching; thus giving more importance to words that correspond to topic during matching.

First of all, we consider that a user expresses the information need by topic only, that is to say that there is no comment in a user's query. For this reason, any query term is considered as a topic in our approach. On the contrary document sentences contain both topic and comment parts. Since users are supposed to be interested by comments about their topic of interest, we hypothesize that the matching model should consider differently topic/query and comment/query matching.

Furthermore, we can assume that matching topics induce that comments are considered relevant information. Thus, the importance of each topic in a document depends not only on its frequency, but also on the number of related comments, i.e. how well the topic is explained in a document. We propose to take the logarithm of this number in order to smooth the influence. On the other hand, some topics may be too specific and thereby linked to few comments. Therefore we introduced the measure of specificity of the topic t Inversed Comment Frequency $ICF(t)$:

$$ICF(t) = \log \frac{\sum_{t_j \in T} CommentCount(t_j)}{CommentCount(t)} \quad (1)$$

where $CommentCount(t)$ is the number of comments related to the topic t in the collection, $T = \{t_j\}_{j=1}^{|T|}$ refers to all topics in the collection, $|T|$ is the total number of topics.

The integration of this proposition in most of IR models is quite simple: a specific document term is considered differently whether it occurs in the topic or the comment part of the sentence. We give the example of the integration into the $BM25F$ retrieval model in the next section.

C. Integration of the Topic-comment Structure into Retrieval Models

We integrated topic-comment structure into $BM25F$ retrieval model. Originally $BM25F$ is an extension of Okapi's $BM25$ to multiple weighted fields in contrast to linear combination of scores for structured documents [27]. $BM25$ is calculated as follows:

$$BM25(d) = \sum_{i=1}^{n} \frac{IDF(q_i) \times TF_d(q_i) \times (k_1 + 1)}{TF_d(q_i) + k_1 \times (1 - b + b \times \frac{|d|}{avgDL})} \quad (2)$$

where q_i are the terms of the query Q, n is the number of query terms, $IDF(q_i)$ is an inverse document frequency of the term q_i, $TF_d(q_i)$ is a term frequency in the document d, $|d|$ is the length of the document d in terms, $avgDL$ is the average document length in the collection, k_1 and b are free parameters. The variable b calibrates the scaling by document length here with $b = 0$ means that there is no length normalization, while $b = 1$ corresponds to the fully scaling [28]. The parameter k_1 determines the document term frequency scaling. Lower values of k_1 tend to a binary model (i.e. without term frequency), while larger values correspond to applying raw term frequency.

$BM25$ model is based on the assumption that term frequencies follow 2-Poisson distribution and for each term the collection is split into two categories: elite and non-elite. As Robertson et al. assert, this relation may be considered from the opposite point of view, namely, the terms of a given document are labeled as elite or non-elite [27]. A term is elite in a document if the document is about the concept denoted by the term. The elite terms refer to the topics of the document. Bag-of-words based approaches presuppose the independence from the position of a term but the boosted probabilities of elite terms. Robertson et al. assumed that for some parts of structured documents the probabilities of the elite terms are boosted even more. Thus, they proposed to assign different weights to the term coming from different document parts:

$$BM25F(d) = \sum_{i=1}^{n} \frac{IDF(q_i) \times TF_d^F(q_i) \times (k_1 + 1)}{TF_d^F(q_i) + k_1 \times (1 - b + b \times \frac{|d|}{avgDL})} \quad (3)$$

$TF_d^F(q_i)$ is a weighted sum of the frequencies of the query term q_i in the document fields:

$$TF_d^F(q_i) = \sum_{f \in d} w_f \times TF_f(q_i) \quad (4)$$

where f are document fields with the corresponding weights w_f and $TF_f(q_i)$ are the frequencies of the query term q_i in the field f.

However, document structure is not uniform and therefore is hard to analyze. In contrast to document fields, topic-comment structure is common for all texts and genres. Thus, we compute document score as follows:

$$score(d) = \sum_{i=1}^{n} \frac{ICF(q_i) \times TC \times (k_1 + 1)}{TC + k_1 \times (1 - b + b \times \frac{len_{topic}(d)}{avgDL_{topic}})} \quad (5)$$

$$TC = tw \times explRate(q_i) f(q_i, T_d) + (1 - tw) \times f(q_i, C_d)$$

$$explRate(q_i) = \log(CommentCount_d(t) + 1)$$

where tw is the topic weight which is the analogue to the field weight in the classical $BM25F$ model, $f(q_i, T_d)$ is q_i's term frequency in the topic parts of the document d, $f(q_i, C_d)$ is the frequency of the term q_i in the comment parts of the document d, $len_{topic}(d)$ is the length of the document d in topics (i.e. the number of topic terms), and $avgDL_{topic}$ is the average document length in the collection in topics, k_1 and b are free parameters, and $CommentCount_d(t)$ refers to the number of comments related to the topic t in the document d. tw is a parameter in the model. It could be assigned or learnt.

Similarly to the classical $BM25$ model, the parameter b determines the scaling by document length but in terms of the number of topics. As in $BM25$, $b = 0$ corresponds to no length normalization, while $b = 1$ indicates the fully scaling. The variable k_1 calibrates topic frequency scaling of a document. As in $BM25$ the weighting parameter tw shows the impact of the topic part of a document into the resulting value.

We introduced the notion of the explanation rate $explRate(q_i)$ showing how well the topic is explained in the document. This notion is similar to the topicality power of a term proposed in [26] which is considered within a document and shows how strong it is explained (i.e. the number of comments it has). The first difference is that we propose to use the logarithm instead a raw sum in order to deal with large numbers. Explanatory power in [26] is viewed as the number of times a term is occurring at a comment regardless the topic within a single document while we are looking for comments to a specific topic. Moreover, in contrast to [26], we consider the collection features by introducing the notion of Inverted Comment Frequency (see Formula 1).

In order to match query terms with topics from documents, after having extracted topic-comment structure, we incrementally extract multi-word expressions based on normalized point-wise mutual information $npmi(x, y)$ [29]:

$$npmi(x, y) = \frac{pmi(x, y)}{-\log[p(x, y)]} pmi(x, y) = \log \frac{p(x, y)}{p(x)p(y)} \quad (6)$$

where $pmi(x, y)$ is the point-wise mutual information of the terms x and y, $p(x, y)$ is the joint probability of x and y, $p(x)$ and $p(y)$ are the probabilities of the terms x and y respectively.

Candidates made of exclusively functional words are rejected as well as candidates containing punctuation marks. We hypothesized that multi-word expression matching should be more important than a single word. Therefore, we integrated the length in terms of tokens of the expression $length(q_i)$ into the final score:

$$score(d) = \sum_{i=1}^{n} \frac{length(q_i) \times ICF(q_i) \times TC \times (k_1 + 1)}{TC + k_1 \times (1 - b + b \times \frac{len_{topic}(d)}{avgDL_{topic}})} \quad (7)$$

IV. Evaluation Framework

The evaluation was performed on two TREC data sets:

- Robust TREC;
- WT10G.

Robust TREC set consists of about 528,000 news articles and 1,904 MB of text of TREC Disk4&5 (except Congressional Record data) and 249 topics with relevance judgments. Robust TREC set is "pure" collections since the documents have almost the same format and there is no spam. WT10G is 10GB subset of the web snapshot and of Internet Archive.

WT10G contains more than 1.6 million of documents. There are 98 topics with relevance judgments. In contrast to Robust, WT10G is a snapshot of the web with real documents in HTML format, some of which are spam.

The system performance was evaluated using several measures implemented in trec_eval[1] software provided by the TREC community for evaluating an ad hoc retrieval run, given the results file and a standard set of judged results. We considered the following evaluation measures:

- MAP (Mean Average Precision) over all queries which is the arithmetic mean of average precision values for individual queries and has been shown to have very good discrimination and stability.
- $NDCG$ (Normalised Discounted Cumulated Gain). Since the gain of each document is discounted at lower ranks, this measures is suitable for re-ranking evaluation.
- $BPREF$ (Binary Preference) computes a preference of whether judged relevant documents have higher rank than judged non-relevant documents. Thus, $BPREF$ does not treat non-assessed documents as non-relevant while MAP does. This is important for large collections where the probability of retrieving non-assessed documents is higher.

We compared our system with two baselines implemented in the Terrier platform [30], namely

- $BM25$;
- $InL2$ weighting model with $Bo2$ query expansion algorithm ($InL2Bo2$).

We used $BM25$ with query term weighting:

$$BM25(d) = \sum_{i=1}^{n} \frac{IDF(q_i) \times TF_d(q_i) \times (k_1 + 1)}{TF_d(q_i) + k_1 \times (1 - b + b \times \frac{|d|}{avgDL})} \times \frac{(k_3 + 1) \times TF_q(q_i)}{(k_3 + TF_q(q_i))} \quad (8)$$

[1] http://trec.nist.gov/trec_eval/

where $TF_q(q_i)$ is the frequency of term q_i in the query q. We used the default values of the model, namely $b = 0.75$, $k_1 = 1.2$, $k_3 = 8$.

$InL2$ is a DFR (divergence from randomness) model based on TF-IDF measure with L2 term frequency normalization [31]. This model is based on the assumption that informative words are relatively more frequent in relevant documents than in others. $InL2$ demonstrates better performance at many recall levels and in average precision than traditional retrieval models such as $BM25$ [32]. $L2$ normalization is less sensitive to document length. According our preliminary study, with the default Terrier's parameters, on the used collections $InL2$ showed better results than Okapi's $BM25$ and Hiemstra's implementation of the language model. $Bo2$ is a pseudo-relevance feedback algorithm for query expansion based on Bose-Einstein statistics and DFR model. On the chosen collections, this method outperformed RM3 model implemented in Indri, a search engine from the Lemur project mainly built on the language modeling information retrieval[2]. RM3 is an Indri's adaptation of Lavrenko and Croft's relevance models [33]. For all method the stemming was performed by Porter algorithm. We parsed the document retrieved by the baseline system by the Stanford POS tagger which also allows sentence chunking [34].

For our model, we used top 20 documents for re-ranking. The re-ranking was performed within blocks of 5 documents. The topic weight was set to $tw = 0.8$. The coefficients $k_1 = 6$ and $b = 0.2$. We considered only unigrams and bigrams. We also excluded the lower order expressions from the query term list if they are parts from a higher order expression. For example, a query $q =$ *safety plastic surgery* is presented as $q = \{q_1, q_2\}$, where $q_1 =$ *safety* and $q_2 =$ *plastic surgery* and the unigrams *plastic* and *surgery* are ignored.

V. Results

Table I provides evaluation results. The differences with the corresponding baselines marked by * are significant at the level $p = 0.05$. According to all evaluation measures for both data sets our method (TC) outperformed the corresponding baselines.

On Robust data set our method $BM25+TC$ showed better results than $BM25$ on 113 queries and it was bellow it on 105 queries. The lower performance was observed for easier queries with the average $NDCG_{avg} = 0.5113$ according to $BM25$ while the better results were obtained for more difficult queries ($NDCG_{avg} = 0.503$). $InL2Bo2 + TC$ excelled the baseline $InL2Bo2$ on 107 queries and it was bellow it on 101 queries. The lower performance was observed for queries with higher values of $NDCG_{avg}$ in average (0.64 according to $InL2Bo2$) while the better results were observed for more difficult queries ($NDCG_{avg} = 0.56$).

On the WT10G $BM25 + TC$ outperformed $BM25$ for 42 queries ($NDCG_{avg} = 0.515$) and it was less efficient for 18 queries ($NDCG_{avg} = 0.55$). $InL2Bo2 + TC$ showed better results than $InL2Bo2$ for 40 queries ($NDCG_{avg} = 0.56$) and it was less efficient for 22 queries ($NDCG_{avg} = 0.628$).

Thus, we can conclude that the approach proposed in this paper is more suitable for difficult queries.

Tables II and III report the detailed statistics of the amelioration/degradation of results for all, very difficult ($MAP(BM25) \leq 0.1$), difficult ($MAP(BM25) \leq 0.25$) and simple ($MAP(BM25) \geq 0.5$) queries for Robust and WT10G collections respectively. These tables also provide evidence that the proposed method improve rather difficult queries especially on the web data set.

Figures 1 and 2 provide the histograms of the NDCG difference between our method and the corresponding baselines on the Robust and WT10G data sets respectively.

In order to evaluate the model stability, we studied the variation of the parameters k_1 and b with the fixed values of the other parameters. Figures 3 and 4 show the influence of b and k_1 respectively on the values of the $NDCG$ on the Robust and WT10G data sets. Here, we presents the results obtained for $BM25$ as a baseline. As previously, we re-ranked 20 documents within blocks of 5 texts. The topic weight was set to $tw = 0.8$. For the variation of k_1 the value of b was set

TABLE I
General results

Collection	Measure	BM25	BM25+TC	InL2Bo2	InL2Bo2+TC
Robust	MAP	0.2365	**0.2386**	0.2801	**0.2884***
	BPREF	0.2462	**0.2472**	0.2782	**0.2863***
	NDCG	0.5079	**0.512***	0.5549	**0.5597***
WT10G	MAP	0.1867	**0.1959***	0.2152	**0.219***
	BPREF	0.1865	**0.1948***	0.2056	**0.2138***
	NDCG	0.4584	**0.4705***	0.4861	**0.4917***

TABLE II
# of improved and worsen queries (Robust)

	All	**Very difficult** $MAP(BM25) \leq 0.1$	**Difficult** $MAP(BM25) \leq 0.25$	**Easy** $MAP(BM25) \geq 0.5$
# of queries	249	10	39	137
$BM25 + TC > BM25$	113	5	20	61
$BM25 + TC < BM25$	105	4	16	58
$InL2Bo2 + TC > InL2Bo2$	107	1	14	61
$InL2Bo2 + TC < InL2Bo2$	101	1	5	65

[2]http://www.lemurproject.org/

TABLE III
OF IMPROVED AND WORSEN QUERIES (WT10G)

	All	**Very difficult** $MAP(BM25) \leq 0.1$	**Difficult** $MAP(BM25) \leq 0.25$	**Easy** $MAP(BM25) \geq 0.5$
# of queries	98	40	71	7
$BM25 + TC > BM25$	42	11	28	4
$BM25 + TC < BM25$	18	4	12	0
$InL2Bo2 + TC > InL2Bo2$	40	16	31	2
$InL2Bo2 + TC < InL2Bo2$	22	9	13	4

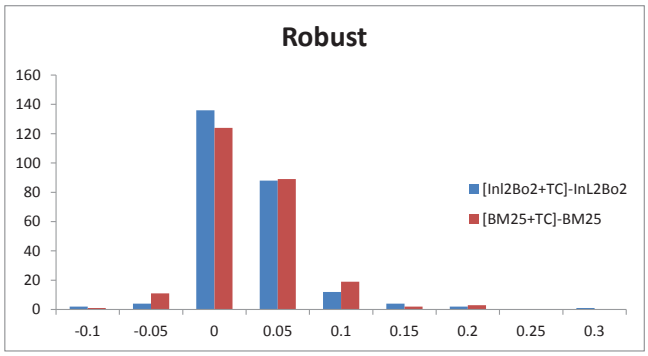

Fig. 1. Histogram of the NDCG difference with the baseline (Robust)

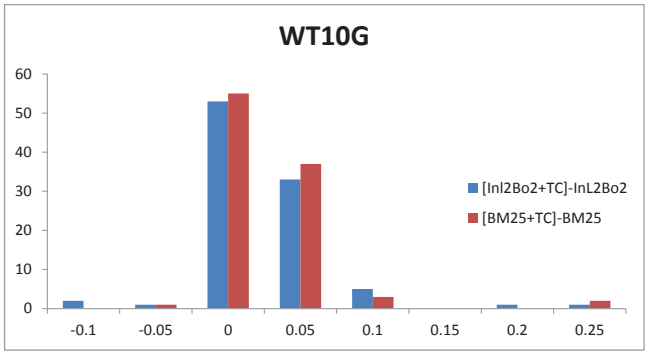

Fig. 2. Histogram of the NDCG difference with the baseline (WT10G)

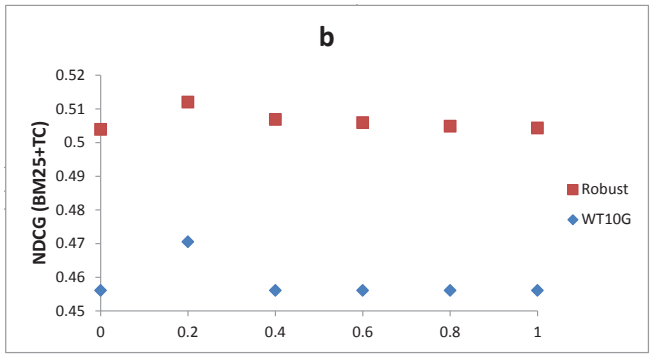

Fig. 3. Influence of the parameter b

to 0.2. k_1 was varied from 2 to 10. For the variation of b the value of k_1 was fixed to 6. We examined the values of b in the inclusive interval $[0, 1]$.

Figure 4 provide evidence that the model is stable regarding the parameter k_1, while Figure 3 indicates that the variation of b influences a lot the re-ranking results. The stability of the proposed method relatively to k_1 means that our model has low sensitivity to term frequency. However, it is very sensitive to the normalization of topic number in a document. The best value of $b = 0.2$ for both collections. It corresponds to low rate of normalization. However, no normalization causes low results. The best value of $b = 0.2$ in our model is lower than the recommended value of $b = 0.75$ in the traditional $BM25$ model. Apparently, it can be explained by the smaller number of topics than the number of terms in a document. b and $k1$ demonstrate the same trends for both collections.

Figure 5 demonstrates the impact of the topic weight tw. Although tw shows stability in general, the trends are different for test collections. For WT10G one can observe that higher topic weights ameliorate results, while for the Robust data set the extreme values provoke small degradation. This could be explained by the fact that the comments are usually much longer than the topics. Thus, the prior probability to find a term within comments is higher than in topics. Higher values of topic weight decrease comment weight. This leads to the lost of documents that just mention relevant information but are not entirely about the subject.

Among 249 queries from the Robust collection 53 queries contained bigrams. For the WT10G this number was equal to 13. We removed these queries in order to measure the performance of the topic-comment approach without bigram extraction. The results are given in Table IV. This table indicates that our approach remains better than the baselines even in case of unigrams.

Let consider a query *piracy* and two examples of documents from the Robust collection.

Example 6:

```
<num> 367
<title> piracy
<desc> What modern instances have there been of old
    fashioned piracy, the boarding or taking control
     of boats?
```

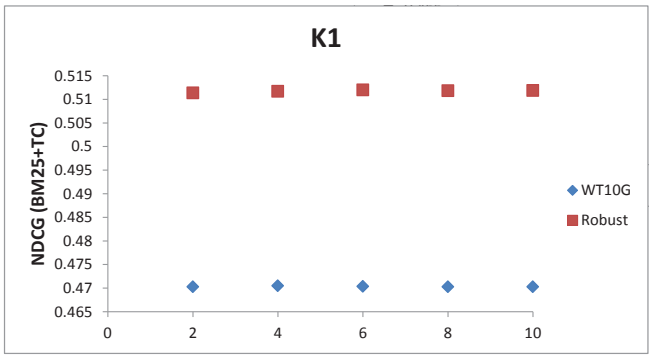

Fig. 4. Influence of the parameter k_1

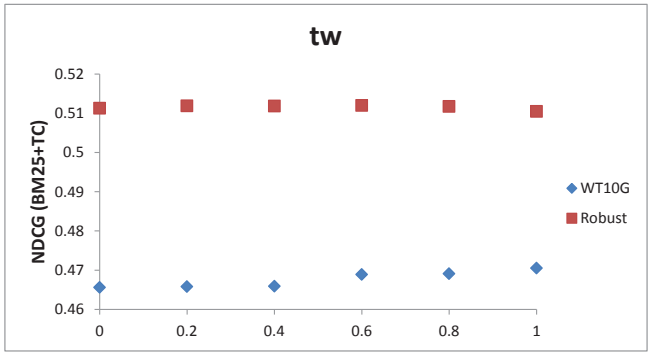

Fig. 5. Influence of the parameter tw

<narr> Documents discussing piracy on any body of water are relevant. Documents discussing the legal taking of ships or their contents by a national authority are non-relevant. Clashes between fishing vessels over fishing are not relevant, unless one vessel is boarded.

Example 7: Document FT923-9880

{FT 03 AUG 92 / Jakarta}$_{topic}$ {sinks plan to combat piracy}$_{comment}$.
{Plans for an international centre to fight the increasing incidence of piracy in south-east Asian waters}$_{topic}$ {have been scuttled}$_{comment}$.
{The International Maritime Bureau (IMB)}$_{topic}$ {had proposed setting up a 24-hour regional centre in Kuala Lumpur to co-ordinate anti-piracy efforts in waters off Malaysia, Singapore, Indonesia and the Philippines}$_{comment}$.
{But Indonesia, in particular,}$_{topic}$ {has objected to what it sees as interference in its affairs}$_{comment}$.
{At a Piracy in South-East Asia conference in Kuala Lumpur, Commodore Sutedjo, director of naval operations and training in the Indonesian navy,}$_{topic}$ {said that as long as piracy occurred within territorial waters, local law enforcement authorities could carry out counter measures more effectively}$_{comment}$.
{There is alarm at the growing frequency and ferocity of the pirate attacks} $_{comment}$.
{More than 40 incidents}$_{topic}$ {have been reported this year in the Strait of Malacca and in the narrow Phillips channel, off Singapore}$_{comment}$.
{Shipowners say most attacks in the area}$_{topic}$ {seem to be carried out by Indonesians who disappear in the labyrinth of Indonesian islands between Singapore and Sumatra}$_{comment}$.
{In one incident pirates}$_{topic}$ { boarded a supertanker carrying 240,000 tons of crude oil in the Phillips channel}$_{comment}$.
{The crew}$_{topic}$ {was tied up and the tanker was left cruising, unpiloted}$_{comment}$.
{Shipowners}$_{topic}$ {have rejected proposals for a toll to keep the region's seas safe}$_{comment}$.
{They}$_{topic}$ {say security is the responsibility of the states themselves}$_{comment}$.
{It was reported last week that Indonesia and Singapore had agreed new measures to combat piracy, including granting each country's marine police and navy the right of hot pursuit}$_{comment}$.

Example 8: Document FBIS4-60337

{BFN [Report by Ahmad 'Izz-al-Din at the Presidential}$_{topic}$.
{Palace--}$_{topic}$ {recorded]}$_{comment}$
{[Excerpt] Prime Minister Rafiq al-Hariri}$_{topic}$ {has denounced Israel's piracy, which contradicts all norms and proves that Israel is not serious about peace}$_{comment}$.
{Prime Minister al-Hariri}$_{topic}$ {denied that there is any hesitation about adopting a stance on the Israeli piracy, noting that Lebanon is studying the possibility of submitting a complaint against this crime}$_{comment}$.
{President Ilyas al-Hirawi and Prime Minister Rafiq al-Hariri}$_{topic}$ {held a meeting this morning during which they discussed the Israeli piracy operation and the measures the government will adopt}$_{comment}$.
{[passage}$_{topic}$ {omitted]}$_{comment}$

Example 7 is talking about pirate attacks and therefore it was judged relevant while the second document (example 8) is rather about politics and thus it was judged irrelevant. Our system assigned higher score to the document FT923-9880 than to the document FBIS4-60337 ($TC(FT923-9880) = 198.98$, $TC(FBIS4-60337) = 160.18$) while $BM25$ ranked these document in the inversed order ($BM25(FT923-9880) = 9.11$, $BM25(FBIS4-60337) = 9.14$) since the term *piracy*

TABLE IV
NDCG VALUES FOR QUERIES WITHOUT BIGRAMS

Collection	BM25	BM25+TC	InL2Bo2	InL2Bo2+TC
Robust	0.4936	0.5002	0.5428	0.5497
WT10G	0.4499	0.4574	0.4881	0.4903

is extremely frequent in the second document. However, it occurs only in the comment part of the second document. In contrast, in the first document it appears both in the topic and the comment parts.

VI. Conclusion

In this paper we proposed a novel approach for document re-ranking in information retrieval based on topic-comment structure of texts, although it can be easily generalized to document retrieval.

We introduced an automatic topic-comment annotation method based on the topic fronting assumption that requires only shallow parsing, namely sentence chunking and POS tagging. The main idea of the proposed method is to split a sentence into two parts by a personal verb.

We integrated topic-comment structure into $BM25F$ retrieval model. Firstly, we hypothesized that the topics should have more weight than the comments. However, the experiment results demonstrated that extreme values of this coefficient (i.e. ignoring topics or comments) decreased the results in average. The possible explanation is that the comments are usually much longer than the topics and therefore the prior probability of a query term to occur within comments is higher. Higher values of topic weight could lead to the lost of documents that just mention relevant information but are not entirely about the subject. In general, the model parameters showed stability, however, the value $b = 0.2$ gives better results. That could be caused by the smaller number of topics with regard to the number of terms in a document.

We evaluated our approach on two TREC data sets. According to all used evaluation measures for both test collections, our method significantly outperformed the strong baselines provided by the Terrier platform. Experiment results allow drawing a conclusion that the approach proposed in this paper is more suitable for difficult queries. Our approach remains better than the baselines even in case of unigrams.

Since our method makes the difference between sentences where the topic and the comment are inversed (as in Examples 3 and 4), we believe that our approach makes sense for question answering and focused IR. In future work we are going to investigate these tracks.

Acknowledgments: The authors would like to thank Ambassade de France en Russie for funding this project (bourse de thèse en cotutelle).

References

[1] K.-F. Wong, D. Song, P. Bruza, and C.-H. Cheng, "Application of aboutness to functional benchmarking in information retrieval," *ACM Trans. Inf. Syst.*, vol. 19, no. 4, pp. 337–370, oct 2001. [Online]. Available: http://doi.acm.org/10.1145/502795.502796

[2] R. Cummins, "A standard document score for information retrieval," in *Proceedings of the 2013 Conference on the Theory of Information Retrieval*, ser. ICTIR '13. New York, NY, USA: ACM, 2013, pp. 24:113–24:116. [Online]. Available: http://doi.acm.org/10.1145/2499178.2499183

[3] J.-Y. Nie, G. Cao, and J. Bai, "Inferential language models for information retrieval," *Transactions on Asian Language Information Processing*, vol. 5, no. 4, pp. 296–322, dec 2006. [Online]. Available: http://doi.acm.org/10.1145/1236181.1236183

[4] D. Bring, *Topic and Comment*. Cambridge University Press, 2011, three entries for: Patrick Colm Hogan (ed.) The Cambridge Encyclopedia of the Language Sciences. Cambridge: Cambridge University Press.

[5] M.A.K.Halliday, *An Introduction to Functional Grammar*, 2nd ed. London: Arnold, 1994.

[6] C. Macdonald, R. McCreadie, R. L. Santos, and I. Ounis, "From puppy to maturity: Experiences in developing terrier," *Proceedings of OSIR at SIGIR*, pp. 60–63, 2012. [Online]. Available: http://opensearchlab.otago.ac.nz/FullProceedings.pdf#page=65

[7] H. Weil, *De l'ordre des mots dans les langues anciennes compares aux langues modernes: question de grammaire gnrale*. Joubert, 1844.

[8] V. Mathesius and J. Vachek, *A Functional Analysis of Present Day English on a General Linguistic Basis*, ser. Janua linguarum : Series practica / Ianua linguarum / Series practica. Mouton, 1975. [Online]. Available: https://books.google.fr/books?id=ZdbLSkaPMJwC

[9] J. M. Hartmann and S. Winkler, "Investigating the role of information structure triggers," *Lingua*, vol. 136, pp. 1–15, 2013.

[10] "Information structure in cross-linguistic corpora: Annotation guidelines for phonology, morphology, syntax, semantics, and information structure," 2007.

[11] P. Cook and F. Bildhauer, "Annotating information structure: The case of topic," in *Proceedings of the Workshop Beyond Semantics. Corpus-based Investigations of Pragmatic and Discourse Phenomena*, 2011, pp. 45–56.

[12] Y. Versley and A. Gastel, "Linguistic tests for discourse relations in the tba-d/z corpus of written german," 2012.

[13] J. Allan, J. Carbonell, G. Doddington, J. Yamron, Y. Yang, J. Allan, B. Archibald, D. Beeferman, A. Berger, R. Brown, I. C. Dragon, G. Doddington, A. Hauptmann, J. Lafferty, V. Lavrenko, X. L. Cmu, S. L. Dragon, P. V. M. Dragon, R. Papka, T. Pierce, J. Ponte, and M. Scudder, "Topic detection and tracking pilot study final report," in *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, 1998, pp. 194–218.

[14] M. Purver, "Topic segmentation," *Spoken language understanding: systems for extracting semantic information from speech*, pp. 291–317, 2011.

[15] B. Hjørland, "Towards a theory of aboutness, subject, topicality, theme, domain, field, content …and relevance," *J. Am. Soc. Inf. Sci.*, vol. 52, no. 9, pp. 774–778, Jul. 2001. [Online]. Available: http://dl.acm.org/citation.cfm?id=380494.380507

[16] N. Suwandaratna and U. Perera, "Discourse marker based topic identification and search results refining," in *Information and Automation for Sustainability (ICIAFs), 2010 5th International Conference on*, Dec 2010, pp. 119–125.

[17] H. Hernault, H. Prendinger, M. Ishizuka *et al.*, "Hilda: a discourse parser using support vector machine classification," *Dialogue & Discourse*, vol. 1, no. 3, 2010.

[18] R. Soricut and D. Marcu, "Sentence level discourse parsing using syntactic and lexical information," in *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, ser. NAACL '03. Association for Computational Linguistics, 2003, pp. 149–156. [Online]. Available: http://dx.doi.org/10.3115/1073445.1073475

[19] L. Carlson and D. Marcu, "Discourse tagging reference manual," *ISI Technical Report ISI-TR-545*, vol. 54, 2001.

[20] C. Lioma, B. Larsen, and W. Lu, "Rhetorical relations for information retrieval," in *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2012, pp. 931–940. [Online]. Available: http://doi.acm.org/10.1145/2348283.2348407

[21] F. Cai, S. Liang, and M. de Rijke, "Personalized document re-ranking based on bayesian probabilistic matrix factorization," in *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, ser. SIGIR '14. New York, NY, USA: ACM, 2014, pp. 835–838. [Online]. Available: http://doi.acm.org/10.1145/2600428.2609453

[22] P. Li, J. Y. Nie, B. Wang, and J. He, "Document re-ranking using partial social tagging," in *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2012 IEEE/WIC/ACM International Conferences on*, vol. 1, Dec 2012, pp. 274–281.

[23] K. Veningston and R. Shanmugalakshmi, "Information retrieval by document re-ranking using term association graph," in *Proceedings of the 2014 International Conference on Interdisciplinary Advances*

in Applied Computing, ser. ICONIAAC '14. New York, NY, USA: ACM, 2014, pp. 21:1–21:8. [Online]. Available: http://doi.acm.org/10.1145/2660859.2660927

[24] S. Chou, J. Zeng, and Z. Dai, "The Application of Semantic Information contained in Relevance feedback in the enhancement of Document Re-Ranking." in *PACIS*, 2014, p. 390. [Online]. Available: http://aisel.aisnet.org/cgi/viewcontent.cgi?article=1387&context=pacis2014

[25] L. Ermakova, "A method for short message contextualization: Experiments at CLEF/INEX," in *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, ser. Lecture Notes in Computer Science, J. Mothe, J. Savoy, J. Kamps, K. Pinel-Sauvagnat, G. Jones, E. SanJuan, L. Cappellato, and N. Ferro, Eds. Springer International Publishing, 2015, vol. 9283, pp. 352–363. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-24027-5_38

[26] A. Bouchachia and R. Mittermeir, "A neural cascade architecture for document retrieval," in *Neural Networks, 2003. Proceedings of the International Joint Conference on*, vol. 3. IEEE, 2003, pp. 1915–1920.

[27] S. Robertson, H. Zaragoza, and M. Taylor, "Simple BM25 extension to multiple weighted fields," in *Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management*, ser. CIKM '04. ACM, 2004, pp. 42–49. [Online]. Available: http://doi.acm.org/10.1145/1031171.1031181

[28] C. D. Manning, P. Raghavan, and H. Schtze, *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[29] G. Bouma, "Normalized (pointwise) mutual information in collocation extraction," in *From Form to Meaning: Processing Texts Automatically, Proceedings of the Biennial GSCL Conference 2009*, vol. Normalized, Tübingen, 2009, pp. 31–40.

[30] I. Ounis, G. Amati, V. Plachouras, B. He, C. Macdonald, and C. Lioma, "Terrier: A High Performance and Scalable Information Retrieval Platform," in *Proceedings of ACM SIGIR'06 Workshop on Open Source Information Retrieval (OSIR 2006)*, 2006.

[31] G. Amati and C. J. Van Rijsbergen, "Probabilistic models of information retrieval based on measuring the divergence from randomness," *ACM Trans. Inf. Syst.*, vol. 20, no. 4, pp. 357–389, Oct. 2002. [Online]. Available: http://doi.acm.org/10.1145/582415.582416

[32] G. Amati, *Probability Models for Information Retrieval Based on Divergence from Randomness: PhD Thesis*. University of Glasgow, 2003.

[33] V. Lavrenko and W. B. Croft, "Relevance based language models," in *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2001, pp. 120–127. [Online]. Available: http://doi.acm.org/10.1145/383952.383972

[34] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky, "The stanford CoreNLP natural language processing toolkit," in *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2014, pp. 55–60. [Online]. Available: http://www.aclweb.org/anthology/P/P14/P14-5010