Open access • Book Chapter • DOI:10.1007/978-3-540-31865-1_10

Filtering for profile-biased multi-document summarization — Source link

Sana Leila Châar, Olivier Ferret, Christian Fluhr

Published on: 21 Mar 2005 - European Conference on Information Retrieval

Topics: Multi-document summarization, User profile and Information extraction

Filtering for Profile-Biased Multi-document Summarization

Sana Leila Châar, Olivier Ferret, Christian Fluhr

Filtering for profile-biased multi-document summarization

Sana Leila Châar, Olivier Ferret, and Christian Fluhr

CEA-LIST/LIC2M
18, route du Panorama - BP6
92265 Fontenay-aux-Roses Cedex
[chaars,ferreto,fluhrc]@zoe.cea.fr

Abstract. In this article, we present an information filtering method that selects from a set of documents their most significant excerpts in relation to a user profile. This method relies on both structured profiles and a topical analysis of documents. The topical analysis is also used for expanding a profile in relation to a particular document by selecting the terms of the document that are closely linked to those of the profile. This expansion is a way for selecting in a more reliable way excerpts that are linked to profiles but also for selecting excerpts that may bring new and interesting information about their topics. This method was implemented by the REDUIT system, which was successfully evaluated for document filtering and passage extraction.

1 Introduction

The need for tools that enable users to face the large amount of documents that are now available in digital form has led the Information Retrieval field to go further than document retrieval. The recent success of the Question/Answering field is representative of this trend. When the query of a user is a factoid question, it is possible to find a short excerpt that contains the expected answer. But when the query is about a topic rather than a fact, an answer can be obtained most of the time only by gathering and putting together several pieces of information coming from several documents. The work we present in this article takes place in this second perspective.

This perspective is also the one of multi-document summarization, which has received great interest during the last years, especially through the DUC (Document Understanding Conference) evaluation [1]. Although query-biased summarization is generally not the focus of the work achieved in this field, this subject was already tackled by work such as [2], [3] or [4]. More recently, a convergence of Question/Answering and multi-document summarization has led to answer-focused summarization ([5], [6]), which was introduced as a task of DUC in 2003.

Our approach, that can be named profile-biased multi-document summarization, is closer to query-biased summarization than to answer-focused one. As we will see in section 3.1, the user profiles we use are not very different from the

TREC topics used in SUMMAC [7] for the evaluation of query-biased summarization. However, our work focuses more specifically on two important points in relation to profiles. First, a profile generally corresponds to a configuration of topics and not only to one large topic. Hence, selecting only the excerpts of documents that refer to all the subtopics of a profile instead of taking the profile as a whole should improve the precision of filtering. Second, profiles are only partial descriptions of the topics they represent. These descriptions can be enriched from the processed documents and adapted to them, which should improve the recall of filtering.

2 Overview

The method we present in this article aims at extracting from a set of documents, for instance the result of a search engine, the text excerpts that match with the information need of a user, expressed in our case by a profile that is structured from a topical viewpoint. This method, which is implemented by the REDUIT system, can be split up into four main steps. First, the input documents are preprocessed, both for selecting and normalizing their content words and segmenting them into topically coherent units that can be compared to the topical units of the profile. The second step, which is a filtering step, is based on the matching between the profile and the topical segments of documents. The result of this matching is first used for discarding the documents without any relation with the profile. Then, it supports the selection of the segments of the remaining documents that match the profile. This selection is also based on the detection of the vocabulary of documents that is closely linked to the profile's one, which is a kind of adaptation of the profile to the documents. Segments whose the selection relies on this extended vocabulary are more likely to contain new information in relation to the profile. Hence, they have a specific status for information filtering: in the following steps, they are processed as the other segments but they are kept separate from them. The third step performs information fusion by detecting and removing redundancies among the selected segments. This operation is first achieved among the segments of a document and then, among the segments coming from all the selected documents. Finally, the fourth step is turned towards users: the selected segments are ranked according to their significance, both from the viewpoint of the profile and the viewpoint of the documents, and they are pruned for limiting the amount of text to read.

3 Profiles

3.1 Structure

In the REDUIT system, users express their needs through profiles. Unlike queries sent to search engines, profiles are used during a long period, which is a reason for asking users to take some time for building them not only as bags of words. More precisely, we chose to structure user profiles according to a topical criterion:

a profile is a set of terms that are grouped into topically homogeneous subsets. This structure aims at improving the precision of filtering by giving a higher score to the documents in which all the topics of a profile are represented. For instance, for a profile about the role of radio during wars (see Table 1), the most relevant documents are those that contain both terms related to war and terms related to radio. A document that only contains terms about war, even if they are numerous, is not likely to be relevant. Only the distinction between the two topics and the dividing of the profile terms according to them can give to a filtering system the ability to discard documents that mainly refer to one of the two topics and to select documents that contain fewer profile terms but terms that are spread in a more balanced way among the two topics. Having such a structure for profiles can improve the precision of filtering but also corresponds to a large part of requests for information coming from users. Those requests are often defined by a configuration of several topics rather by giving only one big topic. The profile of Table 1 is a typical example of this phenomenon.

Table 1. The most significant terms of a profile about the role of radios during war

war subtopic	radio subtopic
guerre (war)	radio (radio)
arme (weapon)	réception (reception)
conflit (conflict)	auditeur (listener)
champ_bataille (battlefield)	capter (to pick up)
cesser_feu (ceasefire)	spot (commercial)
paix (peace)	récepteur (receiver)
agression (aggression)	émetteur (transmitter)
combattant (combatant)	onde_hertzien (Hertzian wave)

As illustrated by Table 1, a topic of a profile is represented by a set of terms. These terms can be mono-terms or compounds. They are normalized by applying the same kind of linguistic preprocessing as the one applied to documents (see section 4.1).

3.2 Topical structuring of profiles

For facilitating the use of the REDUIT system by a user, especially for a new user, we do not impose to him to structure the profiles he defines. In such a case, the user can only give a list of terms and a specific algorithm is applied to structure automatically the profile in a topical way. This algorithm performs the unsupervised clustering of a set of words by relying on a network of lexical co-occurrences. The nodes of this network are the words of the vocabulary of a corpus and the edges between them stand for the co-occurrences found in the corpus between these words. The network we used for this work was built from a 39 million word corpus made of 24 months from the *Le Monde* newspaper.

After a filtering procedure was applied [8] to select the co-occurrences that are likely to be supported by a topical relation, we got a network of 7,200 lemmas and 183,000 co-occurrences.

The clustering algorithm applied to profiles is based on the idea that in such a network, the density of links between the words referring to the same topic is much higher than the density of links between words that are not part of the same topic. Hence, a subtopic in a profile can be identified by the fact that its words form a strongly connected subgraph in the network. The detection of such a subgraph is performed by the following iterative algorithm[1]:

1. selection of the profile words to cluster and building of a topical representation of each of them;
2. building of a similarity matrix for all the selected words of the profile;
3. identification of the most significant subtopic;
4. return to the first step after discarding from the words to cluster those that are part of the new subtopic. The algorithm stops when the number of remaining words is too low (less than 3 words) for building a new subtopic.

The first step exploits the network of lexical co-occurrences for associating to each profile word the words of the network that are the most strongly linked to it in the context of the profile. It also discards the profile words that are not considered as topically significant. As the global algorithm, this step is iterative:

1.1 selection of the words in the network $\{nw_i\}$ that have a minimal number of links, in our case fixed to 4, with profile words;
1.2 selection of the profile words that have supported the selection of a minimal number, fixed to 3, of the $\{nw_i\}$;
1.3 return to step 1.1 with the profile words selected in step 1.2. The process stops when the sets of selected words, both for the profile and the network, are stabilized.

The topical representation associated to each selected word of the profile after this first step is used for evaluating their similarity in step 2. The similarity value of two profile words is the size of the intersection of their topical representations. The similarity vector of a profile word is then filtered for making clustering less sensitive to noise: all values lower than 30% of the maximal value of the vector are set to zero.

The third step is two-fold. First, the seed of a new subtopic is selected. This is the profile word whose the sum of its similarity values with the other profile words is the highest one, that is to say, the word that can be considered as the most central one for the new subtopic. Second, the subtopic is built by aggregating to the initial seed its closest profile words. More precisely, a profile word is associated to the seed if its similarity value with it is the highest of its non-null similarity values with profile words. For extending the new topic, the aggregation step is redone with all the words of the topic as possible targets

[1] The various thresholds hereafter were set experimentally from the CLEF 2003 topics.

and not only its seed. For not introducing noise, the new word must also have a non-null similarity value with a minimal number, fixed to 3 in our case, of words already tied to the new subtopic.

In the more global perspective of the evaluation of the REDUIT system (see section 7), the algorithm for structuring profiles we have presented above was tested on the French version of 200 topics of the CLEF evaluations from 2000 to 2003. Each topic was transformed into a list of content words by applying the same linguistic preprocessing as the one applied to documents (see section 4.1). Only the removal of some "meta-words" (such as *trouver* (to find), *document*, *information* ...) related to the Information Retrieval field was specifically done for CLEF topics [9]. Among these 200 topics, the structuring algorithm found 145 of them with only one topic, 48 with two subtopics and 7 with three subtopics.

4 Filtering

4.1 Preprocessing of documents

The first step of the filtering process is a linguistic and topical preprocessing of the input documents. The goal of this preprocessing step is to represent documents in a same way as profiles to make their comparison easier. The linguistic preprocessing of documents mainly consists in normalizing words of documents and selecting those that are considered as significant from a topical viewpoint. These two tasks are achieved by the LIMA (LIc2m Multilingual Analyzer) tool [9], which performs more precisely the tokenization, the morphological analysis and the Part-of-Speech (POS) tagging of documents. The selection of the topically significant words is based on their POS category: only nouns, verbs and adjectives are kept. The LIMA tool also achieves named entity recognition, that is to say, identifies persons, locations, organizations, dates, numerical values, companies and events.

The topical preprocessing of documents relies on the result of their linguistic preprocessing and aims at segmenting them into topically homogeneous segments. These segments are delimited in our case by the means of the C99 algorithm [10], which is a state-of-the-art linear text segmentation algorithm that only exploits word reiteration. Classically, each segment is represented as a vector of normalized terms according to the Vector Space Model.

4.2 Selection of documents

Principles As the REDUIT system does not go on the assumption that all its input documents are relevant for its current profile, its first task is to discard documents without relation with it. More precisely, the REDUIT system distinguishes three cases for a document and a profile:

- the document globally matches with the profile, even if some of its parts are about topics not in the profile;

– only a part of the document matches with the profile. This one is only a secondary topic of the document;
– the document has no relation with the profile, even locally.

The REDUIT system aims at selecting the documents that come under the first two cases. As we assume that a global match between a document and a profile implies that at least a part of it matches with the profile (see 4.3), the main criterion for selecting a document comes from the second case: a document is selected if at least one of its segments matches with the profile.

Similarity between a profile and a segment As mentioned in section 3.1, the profiles in the REDUIT system are structured from a topical viewpoint to avoid selecting a document or a part of it while it only refers to a part of a profile. Following this principle, a segment of a document can match with a profile only if each subtopic of the profile is represented in the segment. As the size of a document segment is generally equal to the average size of a paragraph, that is not too large, we consider that a subtopic of a profile is represented in a segment when at least one of the terms that defines the subtopic is present in the segment. This criterion may seem not very strict for one topic but it is more significant for a multi-topic profile.

Although compounds are generally less ambiguous than mono-terms, we do not place conditions on the presence of compounds for the identification of topics in segments as we do not want to impose too strict constraints on the way profiles are defined by users. Nevertheless, one can observe that manually built profiles often contain a large number of compounds, which has led us to pay attention to their identification. In order to favor robustness, this identification is not performed in our case by a general terminology extractor but by the set of the following heuristics:

– the words MT_i that are part of a compound CT must occur in a segment in the same order as in CT[2]. The identification of the MT_i directly relies on the linguistic preprocessing of documents;
– an occurrence of CT can not be larger than $1.5 * N$ content words, where N is the number of words in CT. This heuristic takes into syntactical variations such as insertions;
– if CT contains prepositions, they must also be present in its occurrences and their position in relation to the MT_i must be the same as in CT. Moreover, a possible occurrence of CT must not contain any punctuation mark.

A compound CT can also be recognized when only one of its sub-terms ST is identified, which is called *approximate recognition*. Three conditions must be fulfilled for such a recognition:

– ST must contain at least half of the content words of CT;

[2] Of course, this kind of heuristics is less effective for languages where the order of words in compounds is very flexible.

- ST is recognized by fulfilling the three conditions mentioned above for the *strict recognition* of a compound;
- CT must be recognized in a strict way in the document at least one time.

For the identification of a topic in a segment, one of its terms can be recognized in a strict or approximate way.

4.3 Selection of segments

Core segments and extension segments As defined in section 4.2, the selection of a document results from two situations: the document globally matches with the considered profile or only a part of this document matches with the profile. In the second case, as the profile does not correspond to the main topic of the document, there is no reason for selecting other segments than those matching the profile according to the criteria of section 4.2. These segments, whose the matching with the profile is strictly based on the terms of that profile, are called *core segments* and can be viewed as direct instances of the profile in a document.

On the contrary, segments can be selected in the first case according to less reliable criteria because of the global similarity between the document and the profile. More precisely, the selection of a segment can rely on the presence of terms of the document that are considered as linked to those of the profile and not only on terms of the profile. These terms are called *inferred terms*. This is a way to specialize a profile in relation to a document and also a way to detect new trends in relation to the topics of the profile. Hence, when a document globally matches a profile, the criterion for selecting one of its segment is slightly modified: a segment is selected if a term or an inferred term of each subtopic of the profile occurs in the segment. A segment whose the selection is based, at least partly, on inferred terms is called an *extension segment*, as it is more likely to bring new information in relation to the profile.

Selection of inferred terms The detection of a a link between a term of a profile and a term of a document is based on co-occurrences in the document. More precisely, let $\{tp_{Ti}\}$ be the set of terms defining the topic T that are present in the document. $\{td_{Tj}\}$ is the set of terms of the document such that td_{Tj} co-occurs with a term tp_{Ti} in a segment (tp_{Ti} is not necessarily the same term in all these segments). td_{Tj} is considered as an inferred term when this co-occurrence is observed among a significant proportion (1/3 in our experiments) of the segments of the document.

The inferred terms represent a kind of adaptation of profiles in relation to the documents to which they are compared. When a profile is defined manually, the description of its topics tends to be somewhat general. The terms of this description are found in documents but the topics they characterized are also expressed through more specific terms that are not present in the profile and that are useful to identify for improving the results of the filtering process. The

detection of inferred terms is quite similar to the blind relevance feedback used in Information Retrieval.

Table 2 shows the inferred terms extracted from documents of the CLEF 2003 corpus (see section 7) for the profile of Table 1 about the role of radio during wars. Some of these terms, such as *extrémiste, massacrer or défense*, are linked to the topic of war whereas terms such as *station, studio or communiqué* are rather linked to the topic of radio. With the REDUIT system, a user can validate or discard the inferred terms extracted from documents. Moreover, he can dispatch these terms among the topics of the profile or let the system to do it by applying the algorithm described in section 3.2.

Table 2. Example of inferred terms for the profile of Table 1

war subtopic	radio subtopic
extrémiste (extremist)	station (station)
massacrer (to slaughter)	studio (studio)
défense (defense)	appel (call)
exode (exodus)	communiqué (communiqué)
ONU (UN)	programme (program)
génocide (genocide)	BBC (BBC)

Matching of a profile and a document We assume that the global matching of a document with a profile implies that the main topic of the document fits with the topic represented by the profile. Although the problem of identifying the topical structure of texts is far from being solved, the work in the field of automatic summarization exploits an empirical definition of the notion of main topic: the main topic of a text is the topic that is found at the beginning or the end of the text and that covers a significant part of it.

If we transpose this definition in our context, the main topic of a document matches with a profile if the two following conditions are fulfilled:

- the profile must match with the first or the last segment of the document;
- more globally, the segments that match with the profile must represent a significant part of all the segments of the document (1/3 in our experiments).

The first condition relies on the evaluation of the similarity between a segment and a profile presented in section 4.2. The second one is based the extended version of this similarity evaluation (see section 4.3) that takes into account inferred terms.

5 Information fusion

The information filtering performed by the REDUIT system aims at selecting the parts of the filtered documents that are relevant in relation to a profile but also

to detect and to discard redundancies among these selected segments. Hence, the selection stage described in the previous section must be followed by a fusion stage. This fusion, which is first achieved among the segments of a document, then among the segments from several documents, is performed by selecting the segment that conveys in the more representative way the information brought by a set of similar segments.

5.1 Intra-document fusion

As the two types of segments we have distinguished in section 4.3 are complementary, we do not try to detect redundancies between core segments and extension segments. For each kind of segments, the detection of redundancies relies on a the computation of a similarity measure between segments and the comparison of the resulting value to a fixed threshold, T_{fusion}. Classically, we used the cosine measure, which was applied to the segment vectors coming the preprocessing of a document (see section 4.1):

$$sim(S_1, S_2) = \frac{\sum_i freq(t_i, S_1) \cdot freq(t_i, S_2)}{\sqrt{\sum_i freq(t_i, S_1)^2 \cdot \sum_i freq(t_i, S_2)^2}} \quad (1)$$

where $freq(t_i, S_{\{1,2\}})$ is the frequency of the term t_i in the segment $S_{\{1,2\}}$. If the similarity value between two segments is higher than T_{fusion}, they are considered as similar and are supposed to contain roughly the same information. Hence, only one of them can represent the twos for a document. In the opposite case, the two segments are kept.

More globally for a document, the similarity measure (1) is computed for each pair of its core segments and its extension segments respectively. Its segments are then grouped according to their similarity value: each segment is associated to its nearest segment, provided that their similarity value is higher than T_{fusion}. The result of this process is a set of non-overlapping groups of similar segments. If all the selected segments of a document are closely linked to each others, only one group may be formed for one kind of segments.

Then, a representative is selected for each group of similar segments: it is the segment that conveys the largest part of the information that characterizes the group. More specifically, this representative is the segment that contains the largest part of the vocabulary of the group, *i.e.* those of the lemmas of its segments that are shared by at least two segments. This last condition ensures that a segment is selected because its content actually characterizes the group of segments it belongs to and not only because it conveys a large amount of information.

5.2 Inter-document fusion

After the intra-document fusion, each document is represented by two sets (one for core segments and the second one for the extension segments) of segments

that are not similar to each others according to (1). The first step of the inter-document fusion consists in merging all the sets containing the same kind of segments. Two large sets are obtained and in each of them, the algorithm described in section 5.1 for intra-document fusion is applied for detecting redundancies between segments coming from different documents and choosing a representative for each group of similar segments. Finally, the fusion process produces a set of core segments and a set of extension segments.

6 Towards summaries

6.1 Ranking of segments

The REDUIT system is not a fine-grained summarization system which aims at producing short or very short summaries as in the DUC evaluation for instance. Our main objective is rather to design a tool to help users to focus quickly on the document excerpts that are likely to match with their needs. Hence, putting to the front the most relevant of these excerpts is necessary.

As for the fusion step, we chose for this ranking to keep separate the core segments and the extension segments because they represent two complementary aspects of filtering. Each segment is given a relevance score based on its vocabulary. This score takes into account both how well it matches with the profile and how well it is the representative of its group of segments:

$$score(S) = \alpha \cdot \sum_i freq(tps_i, S) + \beta \cdot \sum_i freq(tpa_i, S) + \gamma \cdot \sum_i freq(tcg_i, S) \quad (2)$$

where tps_i is a term of the profile that is recognized in a strict way, tpa_i is a term of the profile that is recognized in an approximate way and tcg_i is one the shared terms of the group of segments whose S is the representative. α, β and γ are modulators which are set in our case[3] in such a way that the stress is put on the similarity with the profile but with a significant place given to the terms linked to it (terms tcg_i). As the size of segments is quite homogeneous, none normalization in relation to this factor was applied to (2).

Finally, segments are ranked in the decreasing order of (2) so that a user can inspect first the segments at top of the list as in a search engine or more radically choose to see only a subset of them by applying a compression ratio.

6.2 Pruning of segments

Although our attention is not focused on the size of summaries, it is quite obvious that the more text a user has to read, the more time he spends for having a view of what could interest him in a set of documents. Hence, the REDUIT system performs a kind of filtering at the segment level. In this case, the basic units

[3] $\alpha = 1.0$, $\beta = 0.75$ and $\gamma = 0,5$

are the sentences of the segment. For selecting coherent units, a sentence always comes with a minimal context made of its N preceding sentences and its N following sentences[4]. Hence, if a segment is not larger than $2N + 1$ sentences, it is selected as a whole. Otherwise, the REDUIT system delimits and selects the groups of adjacent sentences that contain terms of the profile. Two such groups in a segment must be separated by at least $2N + 1$ sentences for not being joined. Moreover, as named entities are considered as especially significant elements, each sentence of a segment that contains at least one of the named entities of the profile is selected with its context.

7 Evaluation

7.1 Methodology

For evaluating the REDUIT system, we adopted an intrinsic method based on the content of documents. This method is an adaptation of the existing evaluations in the summarization field, SUMMAC [7] and DUC [**?**] for English and NTCIR [11] for Japanese, to the characteristics of the REDUIT system, that is to say, a multi-document summarization system for French that is guided by a profile and produces passage-based summaries.

For our evaluation, profiles were, as in the *Ad-Hoc* task of SUMMAC, topics such as those used in the Ad-Hoc task of the TREC evaluation. TREC topics were replaced in our case by topics coming from the CLEF evaluation, which exist for French and many other languages. 14 CLEF topics, that were considered as multi-topic ones, were selected and converted into profiles for the REDUIT system. The profile of Table 1 is an example of these transformed topics. For each one, the CLEF judgment data (*qrels*) give a set of documents from the CLEF collection that were judged as relevant or non-relevant for this topic. For French, the documents are articles from the *Le Monde* newspaper and the *SDA* news agency. Each relevant document (around 20 on average for a topic) was preprocessed to delimit sentences, which are our basic units for summarization, and a manual annotation of the units that fit with the topic of the document was performed to build a "gold standard" for the evaluation of the filtering and the pruning of segments.

The filtering of documents and the filtering of segments were evaluated separately but in both cases, the main objective of our evaluation was to show the interest of taking into account the topical heterogeneousness of profiles. Hence, we compared the results of the filtering with topically structured profiles and its results with the same profiles but without any topical structuring. In this last case, all the terms of each profile were gathered into one topic. In order to make this comparison as objective as possible, the structuring of the 14 test profiles was performed automatically by the structuring algorithm of section 3.2.

[4] In our experiments, N is equal to 1, which means that an excerpt from a segment cannot be smaller than 3 sentences.

Work about summarization evaluation has given rise to several metrics such as the relative utility proposed by [12] or more recently, the ROUGE measure developed in the context of the last DUC conferences [13]. For evaluating REDUIT, we adopted the classical recall/precision measures used in Information Retrieval, as metrics that are more specific to summarization are rather adapted to short summaries. In our context, precision and recall are defined by:

$$precision = \frac{P}{NP + P} \qquad recall = \frac{P}{P + R} \tag{3}$$

where NP is the number of non-relevant units selected by the system, P, the number of relevant units selected by the system and R, the number of relevant units missed by the system. For the filtering of documents, units are documents. For the filtering of segments, they are sentences. Classically, the $F_1 measure$ was used for combining recall and precision. The results presented in the next section are average values of these metrics for the 14 test topics.

7.2 Experimental results

Document filtering Our first evaluation was focused on the ability of the REDUIT system to select documents in relation to a profile. The corpus we relied on was made of the 3780 documents from the CLEF collection for which a relevance judgment against the 14 selected topics was available. It should be note that this corpus can be considered as especially difficult as, according to the pooling procedure used in TREC-like evaluations, it gathers the documents that were considered as the most relevant ones by the search engines that participated to CLEF. Among these 3780 documents, only 320 of them were relevant for the 14 selected topics.

Table 3. Results of the evaluation of document filtering

Filtering method	Recall	Precision	$F_1 measure$
REDUIT (v0)	0.89	0.11	0.21
REDUIT (v1)	0.82	0.44	0.57

In Table 3, REDUIT (v0) is a version of the REDUIT system in which the whole profile is taken as one topic, while REDUIT (v1) is a version that exploits the topical structure of profiles but not the inferred terms. As expected, taking into account the topical heterogeneousness of profiles leads to a significant improvement of precision while recall only decreases slightly. Nevertheless, the global improvement is clear. Precision values are low, which is not surprising: as mentioned above, our corpus is quite difficult and moreover, the filtering of documents with a profile that was defined manually, as TREC-like topics for instance, is known as a difficult task which was given up by the filtering track of TREC [14].

Segment filtering Our second evaluation was dedicated to the ability of the REDUIT system to select the parts of a document that match with a profile. Three versions of the REDUIT system (see Table 4) were tested on the subset of 320 documents that were manually annotated (see section 7.1). REDUIT (v0) and REDUIT (v1) refer to the same versions as in the previous section. Hence, the summaries produced by REDUIT (v0) are made of all the segments of a document that contain terms of the profile, without any constraint on which topic they belong to. Nevertheless, for having comparable results between REDUIT (v0) and REDUIT (v1), the number of profile terms that determines the selection of a segment is the same in the two cases[5]. Table 4 also shows the results of three baseline systems that implement basic strategies that are well-known in the summarization field:

- *baseline 1* always selects the first segment of each document;
- *baseline 2* always selects the first and the last segment of each document;
- *baseline 3* always selects the last segment of each document.

These strategies, that are also used as baselines in the DUC evaluations, rely on the observation that the introduction or the conclusion of a text is frequently comparable to a summary or at least, gather an important part of its content.

Table 4. Results of the evaluation of segment filtering

Filtering method	Recall	Precision	$F_1 measure$
baseline 1	0.56	0.36	0.44
baseline 2	0.68	0.34	0.45
baseline 3	0.11	0.23	0.14
REDUIT (v0)	0.68	0.53	0.6
REDUIT (v1)	0.67	0.65	0.65
REDUIT (v2)	0.82	0.60	0.70

As for document filtering, Table 4 shows the interest of the topical structuring of profiles for segment filtering. The impact on results is the same: precision increases in a significant way while recall decreases slightly, which leads to a clear improvement of global results. Moreover, the results all the versions of the REDUIT system exceed those of all the baseline systems: only the recall of *baseline 2* is comparable to the recall of REDUIT (v0) and REDUIT (v1). This fact can be viewed as an *a posteriori* justification of the criteria defined in section 4.3 for detecting the matching between a profile and a document. The last line of Table 4 corresponds to the most complete version of the REDUIT system, that is to say, a version that also exploits inferred terms. For this evaluation, the dispatching of the inferred terms among the subtopics of the profile was

[5] According to section 4.3, a segment could be selected when one term of the profile is found if this profile has only one topic, which is always the case for REDUIT (v0).

performed manually. In comparison with REDUIT (v1), REDUIT (v2) gets a far better recall with a stable precision, which shows that the kind of adaptation of profiles performed for document filtering is also useful for the selection of segments and more generally, for summarization.

8 Related work

The way we perform query-biased summarization is not radically different from the way it is achieved by Sanderson in [2] for SUMMAC, but we differ from his work by heavily relying on the notion of topic, both for structuring profiles and for delimiting and selecting document excerpts. Sanderson tested a kind of relevance feedback called Local Context Analysis but did not find a positive impact on results, contrary to what we got. One important difference with our inferred words can explain these findings: the topical constraints applied to the selection of inferred words turn out to be quite restrictive, which avoids to introduce too much noise.

The importance of taking into account the topical heterogeneousness of documents was also illustrated for mono-document summarization in [15]. In this case, summarization is not guided by a profile or a query and the topics of a document must be found in an unsupervised way. But this work shows that summaries are less redundant when the selection of sentences is based on their topic as the first criterion than when it relies on a non-topical weighting scheme.

Finally, our work is also related to several evaluations. SUMMAC is the most evident of them as mentioned above but DUC also tested a similar task in 2003: a set of documents had to be summarized given a TDT topic. However, as TDT topics refer to events and not to themes, they are not heterogeneous from a topical viewpoint and the systems developed for this task were mainly focused on taking into account named entities or the semantic links between the topic and the sentences of documents. The HARD track of TREC [16] is also a recent evaluation that pays special attention to profiles. But the focus in this case is put on data related to the context of the query, as the purposes of the user or the kind of documents he is interested in, more than on its topical content.

9 Conclusion and future work

We have presented in this article a method for selecting the most relevant excerpts of a set of documents in relation to a profile. This method puts the stress on two points: taking into account the topical heterogeneousness of profiles can improve the precision of selection; the adaptation of the profile to the input documents can improve its recall. This method was implemented by the REDUIT system and its evaluation showed positive results in favor of its two specificities.

However, several aspects of this work may be improved or extended. For instance, the definition of profiles in REDUIT is done only by giving a set of terms. We are interested in enabling users to define a profile by giving a set of example documents. In this case, a profile would be built by performing the

topical segmentation of the documents and clustering in an unsupervised way the resulting segments for discovering its subtopics. Expanding profiles is also a possible way of improving REDUIT's results. The network of co-occurrences used for structuring profiles was used in [9] for the topical expansion of queries. Such an expansion could be easily adapted to profiles. Finally, a module could be added to REDUIT for detecting more specifically redundancies among the sentences of the pruned segments to produce short or very short summaries.

References

1. Over, P., Yen, J.: An introduction to DUC 2003: Intrinsic evaluation of generic new text summarization systems. In: Document Understanding Conference 2003. (2003)
2. Sanderson, M.: Accurate user directed summarization from existing tools. In: CIKM'98. (1998) 45–51
3. Okumura, M., Mochizuki, H.: Query-biased summarization based on lexical chaining. Computational Intelligence **16** (2000) 578–585
4. Berger, A., Mittal, V.O.: Query-relevant summarization using faqs. In: ACL 2000. (2000) 294–301
5. Wu, H., Radev, D.R., Fan, W.: Towards answer-focused summarization. In: 1^{st} International Conference on Information Technology and Applications. (2002)
6. Mori, T., Nozawa, M., Asada, Y.: Multi-answer-focused multi-document summarization using a question-answering engine. In: COLING 2004. (2004) 439–445
7. Mani, I., House, D., Klein, G., Hirshman, L., Orbst, L., Firmin, T., Chrzanowski, M., Sundheim, B.: The TIPSTER SUMMAC text summarization evaluation. Technical Report MTR 98W0000138, The Mitre Corporation (1998)
8. Ferret, O.: Filtrage thématique d'un réseau de collocations. In: TALN 2003, Batz sur mer, France (2003) 347–352
9. Besançon, R., de Chalendar, G., Ferret, O., Fluhr, C., Mesnard, O., Naets, H.: Concept-based Searching and Merging for Multilingual Information Retrieval: First Experiments at CLEF 2003. In: 4^{th} Workshop of the Cross-Language Evaluation Forum, CLEF 2003. (2004)
10. Choi, F.Y.Y.: Advances in domain independent linear text segmentation. In: NAACL'00. (2000) 26–33
11. Fukusima, T., Okumura, M.: Text summarization challenge: Text summarization evaluation in japan. In: NAACL 2001 Workshop on Automatic Summarization. (2001) 51–59
12. Radev, D.R., Jing, H., Budzikowska, M.: Centroid-based summarization of multiple documents: Sentence extraction, utility-based evaluation, and user studies. In: ANLP/NAACL 2000 Workshop on Automatic Summarization, Seattle, WA (2000)
13. Lin, C.Y., Hovy, E.H.: Automatic evaluation of summaries using n-gram co-occurrence statistics. In: HLT/NAACL 2003, Edmonton, Canada (2003)
14. Hull, D., Robertson, S.: The TREC-8 filtering track final report. In: 8^{th} Text Retrieval Conference (TREC-8). (2000) 35–55
15. Hu, P., He, T., Ji, D.: Chinese text summarization based on thematic area detection. In: ACL-04 Workshop: Text Summarization Branches Out, Barcelona, Spain, Association for Computational Linguistics (2004) 112–119
16. Allan, J.: HARD track overview in trec 2003 - High accuracy retrieval from documents. In: 12^{th} Text Retrieval Conference (TREC-2003). (2004)