Improving Update Summarization by Revisiting the MMR Criterion

Florian Boudin\textsuperscript{1}, Juan-Manuel Torres-Moreno\textsuperscript{1,3} and Marc El-Béze\textsuperscript{1}
\textsuperscript{1} DIRO, Université de Montréal, CP. 6128 succursale Centre-ville, Montréal (Québec), H3C 3J7, Canada
\textsuperscript{2} Laboratoire Informatique d’Avignon BP 1228 84911 Avignon Cedex 09, France
\textsuperscript{3} Ecole Polytechnique de Montréal, CP. 6128 succursale Centre-ville, Montréal (Québec), H3C 3J7, Canada

April 21, 2010

Abstract

This paper describes a method for multi-document update summarization that relies on a double maximization criterion. A Maximal Marginal Relevance like criterion, modified and so called SMMR, is used to select sentences that are close to the topic and at the same time, distant from sentences used in already read documents. Summaries are then generated by assembling the high ranked material and applying some ruled-based linguistic post-processing in order to obtain length reduction and maintain coherency. Through a participation to the Text Analysis Conference (TAC) 2008 evaluation campaign, we have shown that our method achieves promising results.

1 Introduction

Text summarization is the process of automatically creating a compressed version of a given text that provides useful information for the user [10]. Query-oriented summaries focus on a user’s need, and extract the information related to the specified topic given explicitly in the form of a query [8]. On the other hand, generic summaries try to cover as much as possible the information content. Over the past few years, extensive experiments on query-oriented multi-document summarization have been carried out. Extractive summarization produces summaries by choosing a subset of sentences in the original documents. Sentences are then ordered and assembled according to their relevance to generate the summary [22]. This contrasts with abstractive summarization that involves rephrasing information in the text. Although human beings typically produce summaries in an abstractive way, most of the research is on extractive
summarization. This is due to the fact that tools needed to construct semantic representations or generate natural language have not reached a mature stage today. Moreover, existing abstractive summarizers often depend on an extractive component. For example, [25] use a language generation component on top of a multi-document extractive summarizer to produce the final summary. In this paper, we focus on query-oriented multi-document text summarization, where the goal is to produce a summary of multiple documents about a specified topic.

With the ever increasing popularity of news search engines, displaying the information in a more practical and pleasant way is becoming a challenging and important issue. One possible solution is to summarize multiple news so as to propose only one short text instead of raw aggregated headlines. This is, intuitively, a reasonable solution though producing summaries from large collection of documents is a very complicated task. However, as the number of documents increases, facts that are considered as important—and have to appear in the summary—also become more numerous. In this case, a choice must then be made to drop important facts in order to satisfy size constraints. One way to tackle this problem is to remove facts that the user is already aware of. This variant of text summarization is called update summarization. More formally, update summarization is the task of producing summaries while minimizing redundancy with previously read documents (from now on history).

Recently introduced at the Document Understanding Conference (DUC) 2007, update summarization is an emerging summarization task that brings new challenges to sentence ranking algorithms. Indeed, segments have to be selected according to their salience but also to their ability to capture novelty. Existing approaches are derived from state-of-the-art query-oriented multi-document summarizers by the addition of some constraints about redundancy and novelty detection. These include Machine Reading [12], graph-based summarization [17, 31], Maximal Marginal Relevance (MMR) [19], and novelty boosting [4]. The fact that most of them are relying on linguistic resources or tools such as taggers and parsers is a limiting factor for the adaptation to other languages or domains.

In this paper we propose a sentence ranking algorithm inspired by the well-known MMR re-ordering algorithm. Sentences are scored thanks to a double maximization criterion that strives to maximize sentence’s relevance while maximizing non-redundancy with the previously read documents. Our formulation combines word-level similarity measures in an information retrieval approach, ranking sentences by their similarity to the topic and the (inverse) similarity to other sentences in history. We show that our method, although using minimal linguistic resources, can achieve good results among state-of-the-art summarizers. Preliminary results about the sentence re-ranking process were published in [3, 2]. The remainder of this paper is organized as follows. An overview of related work is provided in Section 2. Section 3 presents our three
steps summarization method: pre-processing, sentence ranking and linguistic post-processing. Experimental results are presented in Section 4 followed by discussions and conclusions.

2 Related Work

Introduced by Luhn in the fifties [20], research on automatic summarization can be qualified as a long tradition. In the strategy proposed by Luhn, source sentences are scored for their component word values as determined by $tf \times idf$-type weights. Scored sentences are then ranked and selected from the top until some summary length threshold is reached. Finally, the summary is generated by assembling the selected sentences in original source order. Although fairly simple, this extractive methodology is still used in current approaches. Later on, [9] extended this work by adding simple heuristic features of sentences such as their position in the text or some key phrases indicating the importance of the sentences. As the range of possible features for source characterization widened, choosing appropriate features, feature weights and feature combinations have become a central issue. A natural way to tackle this problem is to consider sentence extraction as a classification task. To this end, several machine learning approaches that uses document-summary pairs have been proposed [15, 28]. Summarization then started gaining more momentum with the SUMMAC evaluation [21], followed by the DUC evaluation conferences.

New tasks have been continuously added to the summarization issue as approaches became more robust and resources grew larger. [1] were amongst the first to tackle the update summarization problem. Their approach, originally developed as a tool to monitor changes in news coverage over time, uses topic detection and tracking techniques to determine which sentences capture usefulness and novelty. The most intuitive way to go about update summarization would be to identify temporal references within documents (dates, elapsed times, temporal expressions, etc.) and to construct a timeline of the events. It is a complex task as temporal references depend on surrounding elements in the discourse but also require an understanding of the ontological and logical foundations of temporal reference construction [13]. Assuming the timeline is constructed, update summaries could be produced by assembling sentences containing the most recent events. However, most recently written material is not necessarily latest facts. This way, focusing the summaries on information that the user is not aware of can be seen as identifying unseen facts. Existing approaches rely exclusively on content-based redundancy removal without recourse to temporal detection. [12] propose a machine reading method to construct knowledge representations from clusters of documents. Sentences that are containing new facts (i.e. that could not be inferred by any document from the history) are selected to generate the summary. A rule-based method using fuzzy coreference cluster graphs was introduced by [31]. This approach can be

2TIPSTER Text Summarization Evaluation Conference (SUMMAC) conducted in May 1998, http://www-nlpir.nist.gov/related_projects/tipster_summac/index.html
applied to various summarization tasks but requires to manually write the sentence ranking scheme. First use a naïve similarity ratio to select sentences that are relevant and dissimilar to sentences from history. On top of this ranking approach, a second method called novelty boosting is used. The latter extends the topic by the unique terms in the cluster, thus biasing the ranking towards maximizing relevance not only with respect to the topic, but also to the novel aspects of the topic in the cluster.

3 Method

In this section we present the details of the proposed text summarization method. As mentioned earlier, our work models sentence ranking as a double maximization criterion. We define \( H \) to represent the previously read set of documents (history), \( Q \) to represent the query and \( s \) the candidate sentence. The following subsections formally define document pre-processing, the sentence scoring method and the summary generation process.

3.1 Pre-processing

The first step is to prepare documents for the ranking process. As we use extractive summarization, documents have to be chunked into cohesive textual segments that will be assembled to produce the summary. The importance of pre-processing is predominant because the selection of segments is based on words they contain \(^3\). The choice was made to split documents into full sentences, in this way obtaining textual segments that are likely to be grammatically correct. Afterwards, sentences are going through several basic normalization steps in order to reduce computational complexity. An example of document pre-processing is given in Table 1. The process is composed by the following steps:

1. **Sentence splitting**: a simple rule-based method is used for sentence splitting. Documents are chunked at the dot, exclamation and question mark signs. Prior to that, ambiguous composed person names (i.e. “George W. Bush”) are detected to reduce segmentation errors.

2. **Sentence filtering**: words are converted to lowercase and cleared up from sloppy punctuation. Words that do not carry meaning such as functional or very common words are removed.

3. **Date normalization**: dates are rewritten and extended with time related words. For example, “december 15, 1982” is replaced by “12/15/1982” and enriched with “_december_1982_”. Standardized dates allow to minimize

\(^3\)The software is available from [http://duc.nist.gov/duc2004/software/](http://duc.nist.gov/duc2004/software/)
the scoring function bias (i.e. considering only one word for one concept instead of three in this example) while enrichment is useful to link facts that were happening at the same period of time (month or year).

4. **Word normalization**: remaining words are replaced by their simplified forms (i.e. inflected forms “go”, “goes”, “went”, “gone”... are replaced by “go”) using a word root database (≈ 88 000 entries). In case of ambiguity, the most frequent word is chosen.

| Original | WASHINGTON _{(1)} A federal judge Monday found President Clinton in civil contempt of court for lying in a deposition about the nature of his sexual relationship with former White House intern Monica S. Lewinsky. Clinton, in a January 1998 deposition in the Paula Jones sexual harassment case, swore that he did not have a sexual relationship with Lewinsky. Clinton later explained that he did not believe he had lied in the case because the type of sex he had with Lewinsky did not fall under the definition of sexual relations used in the case. |
|---|---|
| Splitted | <s0>A federal judge Monday found President Clinton in civil contempt of court for lying in a deposition about the nature of his sexual relationship with former White House intern Monica S. Lewinsky.</s0> <s1>Clinton, in a January 1998 deposition in the Paula Jones sexual harassment case, swore that he did not have a sexual relationship with Lewinsky.</s1> <s2>Clinton later explained that he did not believe he had lied in the case because the type of sex he had with Lewinsky did not fall under the definition of sexual relations used in the case.</s2> |
| Processed | <p0>federal judge monday find president clinton civil contempt court lie deposition nature sex relation former white house intern monica lewinsky</p0> <p1>clinton 01_1998 january_1998 deposition paula jones sex harassment case swear sex relation lewinsky</p1> <p2>clinton late explain believe lie case type sex lewinsky fall define sex relation use case</p2> |

Table 1: Example of pre-processing applied to the document NYT19990412.0403 from cluster D0646A of DUC 2006. News agency name is removed (1); document is segmented into sentences (2); words are normalized (3); punctuation and case are removed (4); dates are standardized end enriched (5).

### 3.2 Ranking

Sentences are scored according to the fact that they contain material satisfying the need formulated in the user’s query. Ranking sentences for query-oriented summarization can be seen as a passage retrieval task in information retrieval.
In this paradigm, sentences sharing most of their vocabulary with the query are likely to be informational for the reader. Each sentence is then scored by computing a combination of two similarity measures with the query. The first similarity measure is the well known cosine \[26\] computed on the sentence and the query vectorial representations in the documents term-space (denoted respectively \(\vec{s}\) and \(\vec{Q}\)). The decision was made not to use the classical \(tf \times idf\) weighting scheme \[27\] because of the difficulty to find similar data and generate pertinent weight lists. The main weakness of cosine and more generally of all similarity measures using words for tokens is that they are relying too much on term normalization. Their performance dramatically decreases with wrongly or non normalized words. That is why we propose a second similarity measure based on the Jaro-Winkler distance \[30\] that can bridge morphologically similar words in order to smooth normalization and misspelling errors. This measure can be classified as an improved edit distance between two word sequences. The Jaro-Winkler distance, denoted \(Jw\), calculates the number of operations required to transform a string into another one. It uses the number of matching characters and transpositions to compute a similarity score between two terms, giving more favourable ratings to terms that match from the beginning. Originally introduced to tackle normalization issues in automatic summarization of chemistry articles \[5\], this distance was extended to compute a similarity measure between a sentence \(s\) and the query \(Q\):

\[
Jw_e(s,Q) = \frac{1}{|Q|} \cdot \sum_{q \in Q} \max_{m \in S'} Jw(q,m)
\]  

(1)

where \(S'\) is the term set of \(s\) in which the terms \(m\) that already have maximized \(Jw(q,m)\) during the previous steps of the summation are removed. The final score is calculated using a linear combination of the two similarity measures. Equation \[2\] shows how to compute the relevance score between a sentence \(s\) and a query \(Q\):

\[
Sim_1(s,Q) = \alpha \cdot \cosine(\vec{s},\vec{Q}) + (1 - \alpha) \cdot Jw_e(s,Q)
\]  

(2)

The Maximal Marginal Relevance (MMR) algorithm \[7\] has been successfully used in query-oriented summarization \[32\]. It strives to reduce redundancy while maintaining query relevance in selected sentences. The summary is constructed incrementally from a list of ranked sentences, at each iteration the sentence which maximizes MMR is chosen:

\[
\text{MMR} = \arg \max_{s \in S} [ \lambda \cdot Sim_1(s,Q) - (1 - \lambda) \cdot \max_{s_j \in E} Sim_2(s, s_j) ]
\]  

(3)

where \(S\) is the set of candidates sentences and \(E\) is the set of selected sentences. \(\lambda\) represents an interpolation coefficient between relevance and redundancy. In
the original formulation, Sim\textsubscript{1} and Sim\textsubscript{2} were computed using the cosine similarity measure. Although this measure has been proven to be efficient, any other similarity measure between sentences remains appropriate.

We propose an interpretation of MMR to tackle the update summarization issue. Unlike previous work such as [19], our approach does not require iterative re-ranking. To remove sentences containing redundant material, the set of selected sentences \( E \) is replaced by the set of sentences in history. In terms of computational complexity, this means that each candidate is compared to all sentences from \( H \). Since Sim\textsubscript{1} and Sim\textsubscript{2} are ranged in \([0, 1]\), they can be seen as probabilities even though they are not. This way, Sim\textsubscript{1} is considered as the probability to be relevant to the topic and Sim\textsubscript{2} as the probability to be redundant with history. We propose to rewrite (3) by adding the constant \((1 - \lambda)\) as (NR stands for Novelty Relevance):

\[
NR = \arg \max_{s \in S} \left[ \lambda \cdot \text{Sim}_1(s, Q) + (1 - \lambda) \cdot (1 - \max_{s_h \in H} \text{Sim}_2(s, s_h)) \right] \tag{4}
\]

This makes more sense because it combines relevance and non-redundancy instead of focusing on redundancy penalization. According to our intuition, we presume that (4) is more or less corresponding to an \(\lor\) (\(\lor\)) combination. But we are obviously looking for a criterion corresponding to \(\land\) (\(\land\)). Since the similarities are independent, we can use the product combination. Sentences are scored thanks to a double maximization criterion in which the best ranked one will be the most relevant to the query \(\land\) the most different to the sentences in \( H \):

\[
\text{SMMR}(s) = \text{Sim}_1(s, Q) \cdot \left(1 - \max_{s_h \in H} \text{Sim}_2(s, s_h)\right) \cdot \mathcal{Nf}(H) \tag{5}
\]

Decreasing parameter \(\lambda\) in (3) with the length of the summary was suggested by [23] and successfully used in the DUC 2005 by [11], thereby emphasizing the relevance at the outset but increasingly prioritizing redundancy removal as the process continues. Similarly, we propose to follow this assumption in SMMR using a function denoted \(\mathcal{Nf}\) that as the amount of data in history increases, prioritizes non-redundancy (\(\mathcal{Nf}(H) \to 0\)). We have defined this parameter function \(\mathcal{Nf}\) as “novelty factor”.

A special breed of redundancy is proliferating in news articles as journalists increasingly rely on the fact that news articles have to be as universally understandable as possible. This means that most of the news articles contain previous facts and/or pointers to previous articles in order for a reader, that does not know anything on the subject, to catch on. This is why we think that a normalized Longest Common Substring (LCS) measure between two sentences
is well adapted to be used as the non-redundancy measure ($Sim_2$). For example, LCS can easily detect sentence rewritings, specially when the sentence is structured around a redundant sub-sentence.

### 3.3 Post-processing

Once sentences are selected to be assembled in the final summary, some linguistic treatments are applied. Indeed, once out of their contexts, discursive forms are considerably decreasing summary’s coherence. For example, two sentences one next to the other in the summary may be in opposition while not dealing with the same subject. Our rule based linguistic post-processing targeted sentence length reduction and coherency maximization. An example of summary post-processing is given in Table 2. The process is composed by the following steps:

1. **Acronym rewriting**: first occurrence of an acronym is replaced by its complete form (acronym and definition); following ones only by their reduced forms. Definitions are automatically mined in the corpus by pattern matching. In case of acronym ambiguity, the most frequent one is selected.

2. **Date and number rewriting**: numbers are reformatted and dates are normalized to the US standard forms (mm/dd/yyyy, mm/yyyy and mm/dd).

3. **Temporal references rewriting**: time tags are used to replace fuzzy temporal references. For example “... the end of next year, ...” with temporal tag 1992_06_02 is replaced by “... the end of 1993, ...”.

4. **Discursive form rewriting**: ambiguous discursive forms are deleted. For example “But, it is ...” is replaced by “It is ...”.

5. Finally, say clauses and parenthesized content are removed and punctuation cleaned.

Sentences are ordered within the summary by original document order and temporal order of documents. Since the acronym rewriting process is dependent to the sentence order and modifies sentence’s lengths, multiple passes are required to generate the final summary. Within summary redundancy is managed by using a simple similarity threshold that prevents duplicate and highly redundant sentences to enter the summary.

---

[^4]: For example, ambiguous say clause “…, he said” is removed.
Last month, U.S. scientists issued a report saying the rate of ice melting in the Arctic is increasing and within a century could lead to summertime ice-free ocean conditions not seen in the area in a million years. The rate of ice melting in the Arctic is increasing and a panel of researchers says it sees no natural process that is likely to change that trend. For example, the white sea ice reflects solar radiation back into space, but as the ice melts the dark water will absorb some of the light, warming and melting more ice. (97 words)

The rate of ice melting in the Arctic is increasing and a panel of researchers says it sees no natural process that is likely to change that trend. The white sea ice reflects solar radiation back into space, but as the ice melts the dark water will absorb some of the light, warming and melting more ice. In 08/2005, US scientists issued a report saying the rate of ice melting in the Arctic is increasing and within a century could lead to summertime ice-free ocean conditions not seen in the area in a million years. (95 words)

Table 2: Example of post-processing treatments applied to the summary produced from cluster D0802A-B of TAC 2008. Dates are standardized (1); sentences are ordered with temporal constraints (2); ambiguous discursive forms are deleted (3).

4 Experiments

The method described in the previous section has been implemented and evaluated by participating to the Text Analysis Conference (TAC) 2008 update summarization track conducted by the National Institute of Standards and Technology (NIST). The following subsections present details of the different experiments.

4.1 The TAC 2008 update track

Piloted in Document Understanding Conference (DUC) 2007, the update summarization task consists in producing a short (100-word) summary of a set of newswire articles, under the assumption that the user has already read a given set of earlier articles. The purpose of each update summary is to inform the reader of new information about a particular topic. The test data-set in TAC 2008 comprises 48 topics. Each topic has a topic statement (examples are given in table 3) and 20 relevant documents which have been divided into two sets: document set A and document set B. Each document set has 10 documents, where all the documents in set A chronologically precede any of the documents in set B. The documents are coming from the AQUAINT-2 collection of news articles.

---

5 More information about the TAC 2008 update track is available at [http://www.nist.gov/tac/](http://www.nist.gov/tac/)

6 [http://duc.nist.gov/](http://duc.nist.gov/)

7 DUC 2007 data was consisting of three temporal document sets A, B and C.
Arctic and Antarctic ice melt (D0802A)
Described the developments and impact of the continuing Arctic and Antarctic ice melts.

Paris Riots (D0819D)
Describe the violent riots occurring in the Paris suburbs beginning October 27, 2005. Include details of the causes and casualties of the riots and government and police responses.

Table 3: Example of topic statements (D0802A and D0819D).

Given a DUC topic and its two document sets (A and B), the task is to create two brief, fluent summaries that contribute to satisfying the information need expressed in the topic statement. The first one is a topic-oriented summary of the document set A while the second one is an update summary of the document set B produced under the assumption that the reader has already read documents in set A.

4.2 Evaluation
All summaries produced by our approach were evaluated both automatically and manually by the NIST. The manual evaluation comprised three scores:

- an Overall Responsiveness score\(^8\) based on both the linguistic quality of the summary and the amount of information in the summary that helps to satisfy the information need expressed in the topic narrative.
- a Linguistic Quality score\(^3\) guided by consideration of the following factors: grammaticality, non-redundancy, referential clarity, focus, structure and coherence.
- a Pyramid \(^{24}\) recall score computed on Summary Content Units (SCUs) annotations. Human annotators select overlapping content in multiple model summaries to construct a pyramid of SCUs.

Most existing automated evaluation methods work by comparing the generated summaries to one or more reference summaries (ideally, produced by humans). In the TAC 2008 evaluation, four human summaries were written for each document set. To evaluate the quality of our generated summaries, several automatic measures were computed:

- ROUGE\(^{18}\) is a \(n\)-gram recall measure calculated between a candidate summary and a set of reference summaries. It is computed as

\(^8\)Integer between 1 (very poor) and 5 (very good).
\(^3\)ROUGE is available at [http://haydn.isi.edu/ROUGE/](http://haydn.isi.edu/ROUGE/)
\[
\text{ROUGE-}(N) = \frac{\sum_{s \in R_{\text{ref}}} \sum_{N\text{-grams} \in s} \text{Co-occurrences}(N\text{-grams})}{\sum_{s \in R_{\text{ref}}} \sum_{N\text{-grams} \in s} \text{Count}(N\text{-grams})} \quad (6)
\]

where \( N \) stands for the length of the \( n \)-gram and \( \text{Co-occurrences}(N\text{-grams}) \) is the maximum number of \( n \)-grams co-occurring in a candidate summary and a set of reference summaries. In our experiments ROUGE-1, ROUGE-2 and ROUGE-SU4 will be computed.

1. **Basic Elements**\(^{10}\) \cite{BasicElements} is similar to ROUGE but uses minimal-length fragments of sensible meaning as units such as “kitchen knife” or “Bank of America”.

In the TAC 2008, NIST received 71 runs from 33 participants for the update summarization task. Each participant submitted up to three runs, ranked by priority. All runs were evaluated automatically (71 runs) but manual evaluations were provided only for runs with priority 1 and 2 (57 runs). In addition, one baseline summarizer was included in the evaluation. It consists in returning all the leading sentences (up to 100 words) in the most recent document. The DUC 2007 update data was used to train our system and to estimate the interpolation coefficient of the similarity measure and the novelty factor. As the DUC 2007 update task was consisting of three temporal documents sets, we have adapted the data set to match the TAC 2008 guideline by removing the third cluster. Parameters for the relevance function and the novelty factor were tuned using this modified data set. The optimal values we have found are \( \alpha = 0.7 \) and \( Nf(H) = 1/c \) with \( c = 1 \) for cluster A (no history) and \( c = 2 \) for cluster B.

### 4.3 Official results

Table 4 shows the results obtained by our submission at the update summarization task of TAC 2008. Our system has achieved good results for Overall Responsiveness and Linguistic Quality, respectively ranked 22\(^{th}\) and 14\(^{th}\) out of 58 submissions, but average ones for automatic evaluations, ranked between the 42\(^{th}\) and 32\(^{th}\) place out of 72 submissions. Giving more confidence to manual evaluation, we can say that our system performed quite well. One surprising result is that our system has obtained high marks in linguistic quality despite

\(^{10}\)Basic Elements is available at [http://haydn.isi.edu/BE/](http://haydn.isi.edu/BE/)
the simplicity of our rule based post-processing.

| Evaluation   | Score  | Rank  |
|--------------|--------|-------|
| Overall Responsiveness | 2.33   | 22/58 |
| Linguistic Quality    | 2.65   | 14/58 |
| Pyramid               | 0.238  | 30/58 |
| ROUGE-1              | 0.33611| 42/72 |
| ROUGE-2              | 0.07450| 38/72 |
| ROUGE-su4            | 0.11581| 32/72 |
| Basic Elements       | 0.04574| 35/72 |

Table 4: Results of manual and automatic evaluations at the TAC 2008 update task.

For a comparative evaluation, Figures 1 and 2 show the results obtained by all the systems participating in the update summarization task at TAC 2008. The baseline consisting of 100-word summaries generated by taking the first sentences in most recent articles is also shown in the two figures. It is worth noting that teams were allowed to submit up to three runs, generally consisting of different parameter configurations. That way, the number of submissions that have obtained better marks than our system may have in fact been produced by a number of systems three times lower. Being more balanced between content and linguistic evaluations, our system always outperforms the widely used lead-based baseline that have been proved to be very challenging [6].

Figure 1: Scatter plot of Linguistic quality and Overall responsiveness for the TAC 2008 update task. Our system (red star) and the baseline (big blue diamond) are highlighted.
Figure 2: Scatter plot of ROUGE-2 and ROUGE-su4 average recall scores for the TAC 2008 update task. Our system (red star) and the baseline (big blue diamond) are highlighted.

Results for separated document sets are presented in Table 5. One can say that evaluation scores are significantly lower for summaries of document sets B but it is worth noting that manual evaluation ranks are significantly better (Overall Responsiveness going from 26th to 16th and Linguistic Quality from 22th to 9th). This shows that, from the linguistic quality point of view, our system is less affected by the increasing difficulty of update summarization than other approaches.

| Evaluation            | Docset A     | Docset B     |
|-----------------------|--------------|--------------|
|                       | score  | rank | score  | rank  |
| Overall Resp.         | 2.417  | 26/58| 2.250  | 16/58 |
| Linguistic Quality    | 2.458  | 22/58| 2.833  | 9/58  |
| Pyramid               | 0.260  | 34/58| 0.215  | 30/58 |
| ROUGE-2               | 0.08125| 36/72| 0.06783| 43/72 |
| ROUGE-su4             | 0.11962| 31/72| 0.11211| 32/72 |

Table 5: Automatic and manual evaluation results for document set A and B.

4.4 Additional results

In these additional experiments, ROUGE scores have been computed using the configuration described in the official guidelines of TAC 2008. To observe the behavior of our method on presence of noisy data, we have added in each cluster a number of random documents taken from different clusters. Since each cluster contains 10 relevant documents, this means a 2/12 (17%), 4/14 (29%) and 10/20 (50%) noise on the data sets. Results on noisy data are given in

---

11 Evaluation guidelines are available at [http://www.nist.gov/tac/tracks/2008/summarization/](http://www.nist.gov/tac/tracks/2008/summarization/)
Table 6. There is no significant performance loss on our method proving that information retrieval approaches are robust for query-oriented summarization.

| Evaluation | 0%       | 17%      | 29%      | 50%      |
|------------|----------|----------|----------|----------|
| ROUGE-1    | 0.33611  | 0.33604  | 0.33585  | 0.33573  |
| ROUGE-2    | 0.07450  | 0.07450  | 0.07450  | 0.07440  |
| ROUGE-su4  | 0.11581  | 0.11579  | 0.11576  | 0.11569  |

Table 6: Comparison of Rouge average recall scores for our system on 17%, 29% and 50% noisy TAC 2008 data.

We also wanted to examine the impact of the novelty factor $N_f$ used in equation (5) on the summaries produced for document sets B. On Figure 3 we observe an improvement of the Rouge scores for all the values greater than zero, obtaining the best results for values comprised between 0.52 and 0.68. The difference with the optimal value found on the training data is minimal but handicap our performance. The size of the adapted DUC 2007 training data was obviously too small (10 topics of 18 documents) to avoid over-fitting problems.

Figure 3: Plot of Rouge average recall scores for docset B summaries in relation to the novelty factor $N_f$ for the TAC 2008 update task.

5 Discussion

The summarizer based on the SMMR sentence scoring algorithm succeeds in identifying most relevant—but containing new facts—sentences from clusters of news articles. The results obtained during the TAC 2008 evaluation prove that our method can achieve good results for both linguistic and content quality. Unlike other approaches, our system does not use large linguistic or knowledge resources which makes it lightweight and easily adaptable to any other language.
or any domain. Computing the whole TAC 2008 test data takes less than five minutes on a 2Ghz dual-core with 1Gb of RAM. As applications that are subject to use update summarization algorithms are gathering tremendous amount of data such as news aggregators, computational complexity is becoming an important feature to take into consideration.

We have observed another interesting result on our submission: automatic and manual evaluations are not often correlated. To illustrate this lack of correlation, the topics that, within our submission, have received the best manual (D0828) and automatic (D0845) scores are compared. Results are shown in Table 7. As we can see, manual and automatic evaluation scores are in total contradiction. Indeed, according to manual evaluations, our best summaries have been generated for the topic D0828 while automatic scores for this topic are poor. Inversely, according to automatic scores, our best topic is D0845 while its manual scores are very poor. By scrutinizing the generated summaries shown in the Table 8, we have identified the reasons of this issue. Redundancy is the main factor for these high ROUGE scores. Units of meaning such as “the ivory-billed woodpecker” are split in an incorrect way, wrongly increasing the number of matching tokens used for computing recall scores. This example proves that using only automatic evaluations is somehow risky.

| Evaluation          | D0828   | D0845   |
|---------------------|---------|---------|
| Overall Responsiveness | 4.0 (1) | 1.5 (35) |
| Linguistic Quality  | 3.5 (6) | 2.0 (29) |
| Pyramid             | 0.324 (1) | 0.215 (38) |
| ROUGE-1             | 0.32993 (26) | 0.39986 (4) |
| ROUGE-2             | 0.06995 (24) | 0.14724 (1) |
| ROUGE-su4           | 0.11299 (28) | 0.18378 (1) |
| Basic Elements      | 0.05562 (18) | 0.05641 (16) |

Table 7: Results of manual and automatic evaluations for topics D0828 et D0845. Ranks obtained by the topic within our submission are shown in parenthesis. The topic ranked in first place contains the summaries that have obtained the best scores in comparison to the other topics of our submission.

6 Conclusions

In this paper we have explained how we had revisited the classical MMR algorithm in order to propose a novel approach to update summarization so called the SMMR. An important aspect of our approach is that it does not requires re-ranking nor linguistic knowledge\(^\text{[12]}\), which makes it a simple and efficient method to tackle the issue of update summarization.

\(^{[12]}\)Our system only uses minimal linguistic resources for post-processing that are easily adaptable to any other language.
### Martha Stewart in Prison

Describe Martha Stewart's experiences while in prison.

| D0828-A | NEW YORK It's check-in day for Martha Stewart. Larry Stewart, who is not related to Martha Stewart, was acquitted of the charges. Q. What will happen to the company, Martha Stewart Living Omnimedia? Stewart spends up to three hours a night writing on a prison typewriter with ribbons purchased at a prison store. Bacanovic and Stewart were both given the option of staying out of prison while they appealed. Martha Stewart has been exercising, reading and making friends in prison, but the food at the minimum-security prison camp in West Virginia is "terrible," the domestic diva's daughter said. |
| D0828-B | Martha Stewart, in a Christmas message posted on her personal Web site, called for sentencing reform and took a swipe at the "bad food" in prison. Since entering federal prison in October, Martha Stewart has tried her hand at ceramics, learned to crochet and become an expert on vending-machine snacks. Martha Stewart, who is about to get out of prison, seems to have undergone a makeover on the cover of the latest Newsweek. One of the tasks ahead of Stewart is to try and spin the goodwill she gained in prison into profits for her Martha Stewart Living Omnimedia Inc. |

### Ivory-billed woodpecker

Describe developments in the rediscovery of the ivory-billed woodpecker, long thought to be extinct.

| D0845-A | The ivory-billed woodpecker, a bird long thought extinct, has been sighted in the swamp forests of eastern Arkansas for the first time in more than 60 years, Cornell University scientists said. "The ivory-billed woodpecker, long suspected to be extinct, has been rediscovered in the 'Big Woods' region of eastern Arkansas," researchers reported in the journal Science to be published. The ivory-billed woodpecker is one of six North American bird species thought to have gone extinct since 1880. The ivory-billed woodpecker, once prized for its plumage and sought by American Indians as magical, was thought to be extinct for years. |
| D0845-B | Recordings of the ivory-billed woodpecker's distinctive double-rap sounds have convinced doubting researchers that the large bird once thought extinct is still living in an east Arkansas swamp. The recordings seem to indicate that there is more than one ivory-billed woodpecker in the area. For half a century, bird-watchers have longed for a glimpse of the ivory-billed woodpecker, a bird long given up for extinct but recently rediscovered in Arkansas. The ivory-billed woodpecker was thought to be extinct until it was spotted in the swamps of southeast Arkansas in 2004. The ivory bill was, or is, the largest North American woodpecker. |

Table 8: Examples of our submission for the topics D0828 and D0845 of TAC 2008.

The novelty factor, characterized in our sentence scoring method by a linear function $N_f(H)$, turns out to be a very important parameter requiring to be tuned in a more judicious manner. Using a linear function that relies on the number of previous clusters instead of the exact amount of text can be hazardous. High redundancy within news articles forces us to believe that the reader
can gain knowledge of only a reduced number of concepts. This is the reason why we think computing the novelty factor by using the concept redundancy is worthy of further work. Recent work by 29 gives some interesting ideas on how to remove redundancy by constructing novel graph-based representations from documents.

It was pointed out that Question Answering and query-oriented summarization have been converging on a common task, the value added by summarization lying in the linguistic quality. We have seen that applying simple ruled-based linguistic treatments to candidate sentences allows to significantly increase the linguistic quality.

Current research works are predominantly focused on the English language. This is why we are currently developing a bilingual evaluation corpus (English and French). Among the others, this point sounds like a promise for further investigation.

Acknowledgments

This work was supported by the Agence Nationale de la Recherche, France, project RPM2. http://labs.sinequa.com/rpm2/

References

[1] J. Allan, R. Gupta, and V. Khandelwal. Temporal summaries of new topics. In Proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval, pages 10–18. ACM New York, NY, USA, 2001.

[2] F. Boudin. Exploration d’approches statistiques pour le résumé automatique de texte. PhD thesis, Université d’Avignon et des pays de Vaucluse, December 2008.

[3] Florian Boudin, Marc El-Bèze, and Juan-Manuel Torres-Moreno. A scalable MMR approach to sentence scoring for multi-document update summarization. In Coling 2008: Companion volume: Posters and Demonstrations, pages 21–24, Manchester, UK, August 2008. Coling 2008 Organizing Committee.

[4] Florian Boudin and Juan-Manuel Torres-Moreno. A Cosine Maximization-Minimization approach for User-Oriented Multi-Document Update Summarization. In Recent Advances in Natural Language Processing (RANLP), pages 81–87, Borovets, Bulgaria, September 2007.

[5] Florian Boudin, Juan-Manuel Torres-Moreno, and Patricia Velázquez-Morales. An Efficient Statistical Approach for Automatic Organic Chemistry Summarization. In Bengt Nordström and Aarne Ranta, editors, 6th International Conference on Natural Language Processing, GoTAL 2008,
[6] R. Brandow, K. Mitze, and L.F. Rau. Automatic condensation of electronic publications by sentence selection. *Information Processing and Management*, 31(5):675–685, 1995.

[7] J. Carbonell and J. Goldstein. The use of MMR, diversity-based re-ranking for reordering documents and producing summaries. In *21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 335–336. ACM Press New York, NY, USA, 1998.

[8] H. Daumé III. *Practical Structured Learning for Natural Language Processing*. PhD thesis, University of Southern California, August 2006.

[9] H. P. Edmundson. New methods in automatic extracting. *Journal of the ACM (JACM)*, 16(2):285, 1969.

[10] G. Erkan and D.R. Radev. LexRank: Graph-based Lexical Centrality as Salience in Text Summarization. *Journal of Artificial Intelligence Research*, 22(2004):457–479, 2004.

[11] B. Hachey, G. Murray, and D. Reitter. The Embra System at DUC 2005: Query-oriented Multi-document Summarization with a Very Large Latent Semantic Space. In *Document Understanding Conference (DUC)*, Vancouver, Canada, October 2005.

[12] A. Hickl, K. Roberts, and F. Lacatusu. LCC’s GISTexter at DUC 2007: Machine Reading for Update Summarization. In *Document Understanding Conference (DUC)*, Rochester, USA, April 2007.

[13] Erhard Hinrichs. Temporal anaphora in discourses of english. *Linguistics and Philosophy*, 9(1):63–82, February 1986.

[14] E. Hovy, C.Y. Lin, L. Zhou, and J. Fukumoto. Automated Summarization Evaluation with Basic Elements. In *Fifth Conference on Language Resources and Evaluation (LREC)*, May 2006.

[15] J. Kupiec, J. Pedersen, and F. Chen. A trainable document summarizer. In *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 68–73. ACM New York, NY, USA, 1995.

[16] Yulia Ledeneva. Effect of preprocessing on extractive summarization with maximal frequent sequences. In *MICAI 2008: Advances in Artificial Intelligence*, volume 5317/2008 of *Lecture Notes in Computer Science*, pages 123–132. Springer Berlin / Heidelberg, 2008.
[17] Wenjie Li, Furu Wei, Qin Lu, and Yanxiang He. PNR2: Ranking sentences with positive and negative reinforcement for query-oriented update summarization. In Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), pages 489–496, Manchester, UK, August 2008. Coling 2008 Organizing Committee.

[18] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In Stan Szpakowicz Marie-Francine Moens, editor, Text Summarization Branches Out: Proceedings of the ACL-04 Workshop, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[19] Z. Lin, T.S. Chua, M.Y. Kan, W.S. Lee, L. Qiu, and S. Ye. NUS at DUC 2007: Using Evolutionary Models of Text. In Document Understanding Conference (DUC), Rochester, USA, April 2007.

[20] H. P. Luhn. The automatic creation of literature abstracts. IBM Journal of Research and Development, 2(2):159–165, 1958.

[21] I. Mani, G. Klein, D. House, L. Hirschman, T. Firmin, and B. Sundheim. SUMMAC: a text summarization evaluation. Natural Language Engineering, 8(01):43–68, 2002.

[22] I. Mani and M.T. Maybury. Advances in Automatic Text Summarization. MIT Press, 1999.

[23] G. Murray, S. Renals, and J. Carletta. Extractive Summarization of Meeting Recordings. In Ninth European Conference on Speech Communication and Technology (Eurospeech), Lisboa, Portugal, September 2005.

[24] Ani Nenkova and Rebecca Passonneau. Evaluating content selection in summarization: The pyramid method. In Daniel Marcu Susan Dumais and Salim Roukos, editors, HLT-NAACL 2004: Main Proceedings, pages 145–152, Boston, Massachusetts, USA, May 2004. Association for Computational Linguistics.

[25] D.R. Radev and K.R. McKeown. Generating natural language summaries from multiple on-line sources. Computational Linguistics, 24(3):469–500, 1998.

[26] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. Communications of the ACM, 18(11):613–620, 1975.

[27] K. Spärck Jones. A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation, 28(1):11–21, 1972.

[28] S. Teufel and M. Moens. Sentence extraction as a classification task. In ACL/EACL workshop on Intelligent and scalable Text summarization, pages 58–65, 1997.
[29] Kapil Thadani and Kathleen McKeown. A framework for decreasing textual redundancy. In Coling 2008, Manchester, UK, August 2008. Coling 2008 Organizing Committee.

[30] W.E. Winkler. The state of record linkage and current research problems. Statistics of Income Division, Internal Revenue Service Publication R, 4, 1999.

[31] René Witte, Ralf Krestel, and Sabine Bergler. Generating Update Summaries for DUC 2007. In Document Understanding Conference (DUC), Rochester, USA, April 2007.

[32] S. Ye, L. Qiu, T.S. Chua, and M.Y. Kan. NUS at DUC 2005: Understanding documents via concept links. In Document Understanding Conference (DUC), Vancouver, Canada, October 2005.