A Maximization-Minimization Approach for Update Text Summarization

Florian Boudin* & Juan Manuel Torres-Moreno*,**

*Laboratoire Informatique d'Avignon, France

**École Polytechnique de Montréal, Canada

Abstract

The work presents an update summarization system that uses a combination of two techniques to generate extractive summaries which focus on new but relevant information. A fast maximization-minimization approach is used to select sentences that are distant from sentences used in already read documents and at the same time close to the topic. On top of this sentence scoring approach, a second method called "Novelty Boosting" is used. The latter extends the topic by the unique terms in the update document cluster, thus biasing the cosine maximization-minimization towards maximizing relevance of a summary sentence not only with respect to the topic, but also to the novel aspects of the topic in the update cluster. Results are based on the DUC 2007 update summarization task.

1 Introduction

Introduced by Luhn (1958) and Rath et al. (1961) in the 50s-60s, research on automatic text summarization can be qualified as having a long tradition. Interest in multi-document summarization started with the on-line publishing and the constant growth of the Internet. Extensive experiments on multi-document summarization have been carried out over the past few years. Most of the strategies to produce summaries are based on extraction methods, which identify salient/relevant textual segments, most often sentences, in documents. Sentences containing the most salient concepts are selected, ordered and assembled according to their relevance to produce summaries (also called extracts) (Mani & Maybury 1999).

Lately emerged from the Document Understanding Conference[1] (DUC) 2007 and then considered as main task during the Textual Analysis Conference[2] (TAC) 2008, update summarization attempts to enhance summarization when more information about the user's knowledge is available. The purpose of each update summary is to inform the reader of new information about a particular topic. In this way, an important issue is introduced:

[1] http://duc.nist.gov/

[2] http://www.nist.gov/tac/

redundancy with previously read documents (also called history) has to be removed from the summary.

A natural way to go about update summarization would to be extract temporal tags (dates, elapsed times, temporal expressions, ...) (Mani & Wilson 2000) or to automatically construct the timeline from documents (Swan & Allan 2000). For the last technique, the well known χ^2 measure (Manning & Schütze 1999) may be used to detect unusual textual segments (words or phrases). These temporal marks could be used to focus extracts on the most recently written facts. However, most recently written facts are not necessarily new facts. (Hickl et al. 2007) propose a Machine Reading (MR) approach 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. However, even though this approach achieves good results (best system at the DUC 2007 update task), it requires very large linguistic resources. A rule-based method using fuzzy coreference cluster graphs was introduced by (Witte et al. 2007). This approach can be applied to various summarization tasks but requires to manually write the sentence ranking scheme. Several strategies relying on post-processing redundancy removal techniques have been suggested. (Lin et al. 2007) have proposed a modified Maximal Marginal Relevance (MMR) (Carbonell & Goldstein 1998) re-ranker during sentence selection, constructing the summary by incrementally re-ranking sentences. More recently, (Boudin et al. 2008) have presented a scalable sentence scoring method derived from MMR. Motivated by the need to detect relevant novelty, candidate sentences are selected according to a combined criterion of query relevance and dissimilarity with previously read sentences.

In this work, we propose a maximization-minimization approach for update summarization. Our method relies on the simple idea that extracts constructed from history can be used to minimize history's redundancy within a candidate summary. The rest of the paper is organized as follows. Section 2 introduces our proposed sentence scoring method. Experimental settings and evaluation details are described in section 3. Results achieved by our approach are presented in section 4, and finally section shows a discussion and conclusion.

2 Method

The underlying idea of our method is that it strives to maximize sentence's salience while minimizing redundancy with the history. In order to do that, we choose a naïve model that relying on a ratio of simple cosine similarity measures. The main advantage of this approach is that *zero knowledge* is required, making the system fully adjustable to any language. We define H to represent the previously read documents (history), q to represent the query (or topic) and s the candidate sentence. The following subsections formally define the maximization-minimization sentence scoring method, the novelty boosting and the surface linguistic post-processings.

2.1 *Query oriented multi-document summarization*

We have started by implementing a *baseline* system for which the task is to produce query-focused summaries from document clusters. Each document is pre-processed: documents are segmented into sentences, sentences are filtered (words which do not carry meaning are removed such as functional words or common words) and words normalized using the well-known Porter algorithm[3] (Porter 1980).

An N-dimensional termspace Γ, where N is the number of different terms found in the corpus, is constructed. Sentences are represented in Γ by a vectors in which each component is the term frequency within the sentence. Sentence scoring for query-oriented summarization can be seen as a passage retrieval task in Information Retrieval (IR). Each sentence s is scored by computing a *cosine* angle measure (Salton et al. 1975) between the sentence and the query vector representations in Γ (denoted respectively $\vec{s}$ and $\vec{q}$) using the well known $tf \times idf$ weighting scheme (Spärck Jones 1972). The score of a sentence s in relation to the query q is:

$$cosine(\vec{s}, \vec{q}) = \frac{\vec{s} \cdot \vec{q}}{\parallel \vec{s} \parallel \parallel \vec{q} \parallel} \tag{1}$$

Sentences coming from different documents are assembled to produce a summary theoretically create redundancy problems for classified document cluster. Moreover, as sentences are all scored by computing a similarity measure with a query, high scored ones are inevitably syntactically related. To tackle this "intra-summary" redundancy issue, a simple but greedy solution is to compare all possible sentence pairs and remove one of two if they are too redundant (i.e., greater than an empirically fixed threshold τ_o). To avoid the quadratic complexity of the process, the redundancy removal is performed during the summary construction by comparing candidate sentences and already selected ones.

[3] i.e., inflected forms such as "*connected*", "*connecting*", "*connection*" ... are replaced by "*connect*".

2.2 *A Maximization-Minimization approach*

In the update summarization task, the main difficulty is that we have to deal with the history's redundancy. The question is how to detect new facts about a particular topic? We propose a naïve model based on the simple assumption that salient sentences are the most relevant to the query (or topic) AND the most different with the history. For efficiency reasons, we choose to represent the history by a cluster's summary (standard query-focused summary) instead of the whole cluster. No difference was found when using only cluster's summaries. This is due to the fact that sentences are selected according to their relevance to a unique topic. Therefore redundant information from the history can only be found in sentences that are also relevant and as a matter of facts used to build the summary up. The score of a sentence s in relation to the query q and the history's summaries $\Pi = \{\vec{p}_1, \vec{p}_2, \ldots, \vec{p}_n\}$ is formally calculated by:

$$\textsc{Max-Min}(s) = \frac{relevance(s,q)}{redundancy(s,\Pi) + 1} \tag{2}$$

$$\begin{array}{lrcl} \text{where} & relevance(s,q) & = & cosine(\vec{s},\vec{q}) \\ \text{and} & redundancy(s,\Pi) & = & \sqrt{\sum_{i=1}^{n} cosine(\vec{s},\vec{p}_i)^2} \end{array}$$

Therefore:

$$\max\left[\textsc{Max-Min}(s)\right] \Longrightarrow \begin{cases} \max\ relevance(\bullet) \\ \min\ redundancy(\bullet) \end{cases} \tag{3}$$

The highest scored sentence s is the most relevant to the topic/query q (i.e., $relevance(s,q) \to 1$) and simultaneously the most different assuming the previous summaries Π (i.e $redundancy(s,\Pi) \to 0$).

2.3 *Novelty boosting*

Detecting novelty is a critical aspect of update summarization. The MAX-MIN sentence scoring method that we propose does not allow the cluster's novelty to enter the summary. We suggest to use the relatedness property of documents within the cluster to expand the information coverage in summaries. In the same way that several previous works in document clustering use a list of high $tf \times idf$ weight terms as topic descriptors (Salton & Yang 1973), we have chosen to represent the most important information of a cluster X by a bag of words bow_X of the highest $tf \times idf$ weight words. The novelty of a cluster of documents A in relation to already processed clusters is the difference of its bag of words bow_A and the intersection of bow_A with all the previous cluster's bags of words:

$$bow_X = bow_X \setminus \bigcup_{i=1}^{n} bow_i \tag{4}$$

This set of terms bow_X is then use to enrich the query q of the cluster X. Selected sentences are then not only focused on the topic but also on "novel" facts.

2.4 *Summary generation*

The summary is constructed by arranging the most highly scored material until a word limit is reached. As it is unlikely that the assembled sentences exactly reach the limit size, extra sentences are considered and the best subset, according to relevance scores, is selected to generate the summary as close as possible to the word limit. Two partial orders are used for sorting sentences within the summary: sentence order within the source document, temporal order of documents within the cluster. Rule based linguistic post-processings are applied to each candidate sentence in order to reduce length and maximize coherency. The process included the following steps:

- Acronym rewriting: the first occurrence is replaced by the full form (acronym and definition), next ones by the reduced forms (acronym only). Definitions are automatically detected in the corpus using patterns.
- Dates and number rewriting: numbers are reformatted and dates are normalized to the US standard forms (MM/DD/YYYY, MM/YYYY and MM/DD).
- Finally, "say clauses"[4] and parenthesised content are removed and punctuation cleaned.

3 Experimental settings

3.1 *Description of the* DUC *2007 pilot task*

The DUC 2007 update task goal is to produce short (~100 words) multi-document update summaries of newswire articles under the assumption that the user has already read a set of earlier articles. The purpose of each update summary will be to inform the reader of new information about a particular topic. Given a DUC topic and its three document clusters: A, B and C, the task is to create from the documents three brief, fluent

[4] As *He said* ... *She says* ... etc.

summaries that contribute to satisfying the information need expressed in the topic statement.

1. A summary of documents in cluster A.
2. An update summary of documents in B, under the assumption that the reader has already read documents in cluster A.
3. An update summary of documents in C, under the assumption that the reader has already read documents in clusters A and B.

Within a topic, the document clusters must be processed in chronological order. Therefore we cannot look at documents in cluster B or C when generating the summary for cluster A, and we cannot look at the documents in cluster C when generating the summary for cluster B. However, the documents within a cluster can be processed in any order. The corpus is composed of 10 topics, with 25 documents per topic. There is approximately 10 documents in cluster A, 8 in cluster B, and 7 in cluster C.

3.2 *Evaluation*

The method described in the previous section (c.f. section 2) has been implemented and evaluated by participating to the DUC 2007 pilot task. Both manual and semi-automatic evaluation were conducted on the summaries produced by our system. A score of Content Quality (Content Responsiveness), ranging from 1 (very poor) to 5 (very good), is manually granted to each summary according to the amount of information that actually helps to satisfy the information need expressed in the topic statement. Most existing automated evaluation methods work by comparing the generated summaries to one or more reference summaries (ideally, produced by humans). To evaluate the quality of our summaries, we choose to use the ROUGE[5] (Recall-Oriented Understudy for Gisting Evaluation) (Lin 2004) evaluation toolkit. ROUGE measures have been found to be well correlated with human judgments. ROUGE-N is an N-gram recall measure calculated between a candidate summary and a set of reference summaries. In our experiments three recall measures will be computed: ROUGE-1, ROUGE-2 and ROUGE-SU4. ROUGE-SU4 is an extended version of ROUGE-2 that allows word-level gaps of maximum length 4 between the bigram tokens. For this evaluation, four reference summaries were manually produced for each cluster. ROUGE has been run with the following parameters:

```
ROUGE-1.5.5.pl -n 2 -x -m -2 4 -u -c 95 -r 1000 -f A -p 0.5 -t 0 -d
```

`-n 2`	compute ROUGE-1 and ROUGE-2
`-x`	do not calculate ROUGE-L

[5] ROUGE is available at `http://haydn.isi.edu/ROUGE`

`-m`	apply Porter stemmer on both models and peers
`-2 4`	compute Skip Bigram with a maximum skip distance of 4
`-u`	include unigram in Skip Bigram (ROUGE-S)
`-c 95`	use 95% confidence interval
`-r 1000`	bootstrap resample 1000 times
`-f A`	scores are averaged over multiple models
`-p 0.5`	compute F-measure with alpha = 0.5
`-t 0`	use model unit as the counting unit
`-d`	print per-evaluation scores

Rouge considers various length fragments to be equally important, a factor that rewards low-informativeness fragments, such as "*of the*", unfairly to relative high-informativeness ones, such as person names. Two evaluation measures have emerged to address these problems. (Hovy et al. 2006) recently proposed to use very small units of content, called *Basic Elements* (BE), expressed as triples (head | modifier | relation). For example, the phrase "*sanction against Russia*" produces the BE (sanction | Russia | against). The *Basic Elements* evaluation consists in computing recall measures similarly to Rouge but using BE instead of N-grams. The idea of using different fragments was taken further by (Nenkova et al. 2005), who named fragments Summary Content Units (SCUs), and deployed them in the *Pyramid* method. The *Pyramid* method uses multiple human summaries to create a gold-standard and exploits the frequency of information in the human summaries to assign importance to different facts. *Basic Elements* and *Pyramid* evaluations conducted during the DUC evaluation are shown in our experiments.

4 Experiments

This section presents the results obtained by our system (the system's id is 47) at the DUC 2007 update task in comparison to the 23 other participants. No training corpus was, at the time of submission, available and there was, as far as we know, no equivalent corpora for training systems. Only manual evaluation of the output summaries was possible. This explains why the parameters used for the system submission are not the optimal ones. The following parameters have been used for the final evaluation: Bag of words size: 15, Redundancy threshold: $\tau_o = 0.4$, minimal sentence length: 5. To have an idea of the relative success of our methods among other summarization systems, we have compared our scores with other participants scores. Table 1 sums up the results achieved by our system at the DUC 2007 update task.

Our system achieves very promising results in both semi-automatic and manual evaluations. Among the 24 participants, our system ranks between

Evaluation	**Score**	**Rank**	**Min**	**Max**
Avg. Content Responsiveness	2.63	7/24	1.67	2.97
ROUGE-1	0.35744	4/24	0.26170	0.37668
ROUGE-2	0.09387	4/24	0.03638	0.11189
ROUGE-SU4	0.13052	5/24	0.07440	0.14306
Basic Elements	0.05458	4/24	0.01775	0.07219
Pyramids	0.27267	5/24	0.07404	0.34031

Table 1: *Official results of manual and semi-automatic evaluations for our system (id is 47) at the DUC 2007 update task*

4^{th} and 5^{th} in semi-automatic measures and the 7^{th} in content responsiveness. The average content responsiveness score obtained by our system is 2.63, which is above the mean (2.32 with standard deviation of 0.35). An example of the best scored topic of our submission (D0726) is presented in Table 2. It contains the three summaries and the full topic statement.

From a reading of the generated summaries we can see clearly that sentence transitions are clumsy. Summary's fluency suffers from a lack of advanced linguistic treatments. Indeed, the high number of repeated person names (textual units referring to "*Al Gore*" are occurring in all sentences) show the limitations of our linguistic post-processing rule-based method. Anaphora's generation is one possible solution to the fluency issue. Replacing person names by personal pronouns allows, as well as to improve readability, to increase the compression rate. Semi-automatic and manual evaluation scores for the best and the worst topic of our submission are shown in Table 3.

Semi-automatic and manual evaluations are, in this example, highly correlated. Only the *Pyramids* score is not in agreement with the manual content quality score. Since *Pyramids* scores are computed according to the number of semantic units, the high number of person names rehearsals in summaries of the topic D0726 can explain the low value. Best few scores for each method are often statistically indistinguishable from the best system in the official evaluations considering the 95% confidence interval. However, enumerate systems that performs significantly better and lower than our approach can be done by studying confidence intervals from semi-automatic evaluations. Table 4 shows these results for our system.

Most of the scores achieved by our approach are better than the other systems. In the DUC 2007 pilot task, only one system can be classified as significantly better than our approach. This system was proposed by (Hickl et al. 2007) and is based on a Machine Reading (MR) approach that constructs knowledge representations from clusters of documents. It is worth noting that our approach is simple, efficient (complexity is $O(n)$) and

Al Gore's 2000 Presidential campaign *Give the highlights of Al Gore's 2000 Presidential campaign from the time he decided to run for president until the votes were counted.*	
D0726F-A	Vice President Al Gore's 2000 campaign has appointed a campaign pro with local Washington connections as its political director. Al Gore, criticized for not having enough women in his inner circle, has hired a veteran female strategist to be his deputy campaign manager for his 2000 presidential bid. Al Gore will take his first formal step toward running for president in 2000 by notifying the Federal Election Commission that he has formed a campaign organization, aides to the vice president said. Al Gore took his presidential campaign to a living room that helped launch Carter and Clinton into the White House.
D0726F-B	Patrick Kennedy, D-R.I., endorsed Vice President Al Gore for the Democratic presidential nomination in 2000. Al Gore named a veteran of the Clinton-Gore presidential campaigns to be his campaign press secretary. Bradley retired from the Senate in 1996, briefly mulled an independent run for president, then spent time lecturing at Stanford University in California before deciding to challenge Gore for the Democratic presidential nomination. Klain was criticized by some Gore allies after President Clinton called a reporter for The New York Times and said Gore needed to loosen up on the campaign trail. Bill Bradley of New Jersey, Gore's sole competitor.
D0726F-C	After hearing that Stamford-native Lieberman had been chosen as Al Gore's running mate, Marsha Greenberg decided to knit him a gift. Vice President Al Gore, who continues to reshuffle his struggling presidential campaign, has selected Donna Brazile to be his new campaign manager, officials said. Al Gore declared "a new day" in his presidential bid with a symbolic homecoming and the opening of a new campaign headquarters far from the constant political intrigue and daily odds-making of Washington. Coelho, Brazile and Carter Eskew, the media consultant hired to help develop Gore's campaign message, are already working out of the Nashville office.

Table 2: *Example of topic (D0726F) coming from our submission. Some post-processing errors may appear showing the limitations of our rule-based method*

Evaluation	**D0726**	**D0743**
Avg. Content Responsiveness	**3.66**	1.66
ROUGE-1	**0.38714**	0.26353
ROUGE-2	**0.11246**	0.05346
ROUGE-SU4	**0.14594**	0.08103
Basic Elements	**0.07491**	0.04282
Pyramids	0.15583	**0.18920**

Table 3: *Results of manual and semi-automatic evaluations for the topics D0726 and D0743. The first one is the best summary of our submission while the second one is the worst*

Evaluation	**Score**	**Lower**	**Upper**	**nb. >**	**nb. <**
ROUGE-1	0,35744	0,01110	0,01112	3	15
ROUGE-2	0,09387	0,00788	0,00815	1	15
ROUGE-SU4	0,13052	0,00721	0,00750	1	16
Basic Elements	0,05458	0,00715	0,00777	1	14

Table 4: *Semi-automatic evaluations for our system at the DUC 2007 update task with lower/upper limits for each score and the number of significantly better (nb. >) and lower (nb. <) systems*

do not uses any linguistic or knowledge resources. Computing the whole DUC 2007 update corpus takes less than a minute on a 2.2Ghz dual-core with 1Gb of RAM running Mac OSX 10.5.4.

5 Discussion and conclusion

We have presented a maximization-minimization approach for producing query-oriented update text summaries. Sentences are scored following a double criterion: maximizing relevance and minimizing redundancy with the history. It is essentially an Information Retrieval (IR) approach, ranking sentences by their similarity to the topic and the dissimilarity to other sentences in a summary. We have introduced a novelty boosting technique that detect important terms in a cluster that have not been mentioned in history, and are thus considered as "novel". We have evaluated our approach by participating in the DUC 2007 update task where our system has done very well: between 4^{th} and 5^{th} in semi-automatic measures and at the 7^{th} in the manual content responsiveness evaluation among 24 participants.

Several important directions for future research are suggested by the results of our experiments. The first one is the impact of novelty boosting on sentence selection. Indeed, extracted terms that we consider as cluster's novelty, are not necessarily topical. In fact, most information in a cluster of documents might be completely irrelevant for the topic. Novelty boosting is hence a way to make the summary more generic, including information that is readily available in the cluster rather than information that is of interest to the user. Bag-of-words extraction have to be redefined so that it includes novel information that has not appeared in history AND still relevant to the topic.

It was pointed out at the DUC 2007 Workshop that Question Answering (QA) and Query-oriented Summarization have been converging on a common task. The value added by summarization lies in the linguistic quality. Approaches mixing IR techniques are well suited for Query-oriented Summarization but they require intensive work to make the summary fluent and coherent. It may seem conflicting for a compression task but adding words in summaries is one way to smooth sentence sequencing and as a result, improve readability. Among the others, this is a point that we think is worthy of further investigation.

Acknowledgments. The authors would like to thank the Journal of Interesting Negative Results (JINR) reviewers for their useful comments. This work was supported by the *Agence Nationale de la Recherche*, France, project RPM2[6].

REFERENCES

Boudin, F., M. El-Bèze & J.M. Torres-Moreno. 2008. "A Scalable MMR Approach to Sentence Scoring for Multi-Document Update Summarization". *Proceedings of the 22nd International Conference on Computational Linguistics (COLING'08)*, Companion volume: Posters and Demonstrations, 21-24. Manchester, U.K.

Carbonell, J. & J. Goldstein. 1998. "The Use of MMR, Diversity-based Reranking for Reordering Documents and Producing Summaries". *Proceedings of the 21st Annual International ACM SIGIR conference on Research and Development in Information Retrieval*, 335-336. Melbourne, Australia.

Hickl, A., K. Roberts & F. Lacatusu. 2007. "LCC's GISTexter at DU 2007: Machine Reading for Update Summarization". *Document Understanding Conference (DUC-2007)*, 56-63. Rochester, USA.

Hovy, E., C.Y. Lin, L. Zhou & J. Fukumoto. 2006. "Automated Summarization Evaluation with Basic Elements". *Proceedings of the Fifth Conference on Language Resources and Evaluation (LREC)*, 899-902. Genoa, Italy.

Lin, Z., T.S. Chua, M.Y. Kan, W.S. Lee, L. Qiu & S. Ye. 2007. "NUS at DUC 2007: Using Evolutionary Models of Text". *Document Understanding Conference (DUC-2007)*, 78-85. Rochester, U.S.A.

Lin, C.Y. 2004. "Rouge: A Package for Automatic Evaluation of Summaries". *Proceedings of the Workshop on Text Summarization Branches Out*, 25-26. Barcelona, Spain.

Luhn, H.P. 1958. "The Automatic Creation of Literature Abstracts". *IBM Journal of Research and Development* 2:2.159-165.

Mani, I. & M.T. Maybury. 1999. "Advances in Automatic Text Summarization". Cambridge, Mass.: *MIT Press*.

Mani, I. & G. Wilson. 2000. "Robust Temporal Processing of News". *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL'00)*, 69-76. Hong Kong, China.

Manning, C.D. & H. Schütze. 1999. "Foundations of Statistical Natural Language Processing". New York, N.Y.: *MIT Press*.

[6] `http://labs.sinequa.com/rpm2`

Nenkova, A. & R. Passonneau. 2004. "Evaluating Content Selection in Summarization: The Pyramid Method". *Proceedings of the Human Language Technology conference / North American chapter of the Association for Computational Linguistics annual meeting* (*HLT-NAACL*). 145-152. Boston, U.S.A.

Porter, M. 1980. "An Algorithm for Suffix Stripping". *Program* 14:3.130-137.

Rath, G., A. Resnick & T. Savage. 1961. "The Formation of Abstracts by the Selection of Sentences". *American Documentation* 12:2.139-143.

Salton, G., A. Wong & C.S. Yang. 1975. "A Vector Space Model for Automatic Indexing". *Communications of the ACM* 18:11.613-620.

Salton, G. & C.S. Yang. 1973. "On The Specication Of Term Values In Automatic Indexing". *Journal of Documentation* 29:4.351-372.

Spärck Jones, Karen. 1972. "A Statistical Interpretation of Term Specificity and its Application in Retrieval". *Journal of Documentation* 28:1.11-21.

Swan, R. & J. Allan. 2000. "Automatic Generation of Overview Timelines". *Proceedings of the 23rd Annual International ACM SIGIR conference on Research and Development in Information Retrieval*, 49-56. Athens, Greece.

Witte, R., R. Krestel & S. Bergler. 2007. "Generating Update Summaries for DUC 2007". *Document Understanding Conference* (*DUC-2007*), 110-113. Rochester, USA.