A Graph-based Approach to Cross-language Multi-document Summarization

Florian Boudin, Stéphane Huet, and Juan-Manuel Torres-Moreno

***Abstract*—Cross-language summarization is the task of generating a summary in a language different from the language of the source documents. In this paper, we propose a graph-based approach to multi-document summarization that integrates machine translation quality scores in the sentence extraction process. We evaluate our method on a manually translated subset of the DUC 2004 evaluation campaign. Results indicate that our approach improves the readability of the generated summaries without degrading their informativity.**

***Index Terms*—Graph-based approach, cross-language multi-document summarization.**

I. Introduction

THE rapid growth and online availability of information in numerous languages have made cross-language information retrieval and extraction tasks a highly relevant field of research. Cross-language document summarization aims at providing a quick access to information expressed in one or more languages. More precisely, this task consists in producing a summary in one language different from the language of the source documents. In this study, we focus on English to French multi-document summarization. The primary motivation is to allow French readers to access the ever increasing amount of news available through English news sources.

Recent years have shown an increased amount of interest in applying graph theoretic models to Natural Language Processing (NLP) [1]. Graphs are natural ways to encode information for NLP. Entities can be naturally represented as nodes and relations between them can be represented as edges. Graph-based representations of linguistic units as diverse as words, sentences and documents give rise to efficient solutions in a variety of tasks ranging from part-of-speech tagging to information extraction, and sentiment analysis. Here, we apply a graph-based ranking algorithm to multi-document summarization.

A straightforward idea for cross-language summarization is to translate the summary from one language to the other. However, this approach does not work well because of the errors committed by Machine Translation (MT) systems. Indeed, translated sentences can be disfluent or difficult to understand. Instead, we propose to consider the translation quality of the French sentences in the sentence selection process. More precisely, we use a supervised learning approach to predict MT quality scores and integrate these scores during the graph construction.

This paper is organized as follows. We first briefly review the previous work, followed by a description of the method we propose. Next, we present our experiments and results. Lastly, we conclude with a discussion and directions for further work.

Manuscript received November 9, 2010. Manuscript accepted for publication January 15, 2011.

Florian Boudin and Stéphane Huet are with Université d'Avignon, France. Juan-Manuel Torres-Moreno is with Université d'Avignon, France; École Polytechnique de Montréal, Canada; Universidad Nacional Autónoma de México, Mexico (e-mail: firstname.lastname@univ-avignon.fr).

II. Related Work

A. Predicting Machine Translation Quality

Machine translation is a natural component for cross-language document summarization. However, as an automatic process, MT systems are prone to generate errors and thus to mislead summarization. These errors can either introduce wrong information with respect to the source-language documents to summarize or make sentences disfluent and difficult to understand. In order to alleviate these effects, it is relevant to take into account a score that assesses the translation quality and that can be used to filter out incorrect translations during summarization.

Predicting quality translation, referred to as confidence estimation in the MT domain, has first been viewed as a binary classification problem to distinguish good translations from bad ones [4]. More recent studies have been done to estimate a continuous quality score at the word level [19] or at the sentence level [19], [20]. In this paper, we choose to resort to sentence-level quality scores that are more easily integrated into the summarization sentence extraction process.

Various classifiers have been used to estimate translation quality. Statistic models are trained on a set of translations manually labeled as correct or incorrect [17], [20] or tagged through automatic metrics like word error rate [4], NIST [4], [20] or BLEU scores [19]. Various features are extracted to compute quality values: linguistic features depending or not on resources like parsers or Wordnet, similarity features between the source sentence and the target sentence and some internal features of the MT system, such as the alternative

translation per source words or the phrase scores of n-best list of translation candidates.

B. Graph-Based Summarization

Extensive experiments on multi-document summarization have been carried out over the past few years, especially through the DUC (Document Understanding Conference) evaluations.[1] Most of the proposed approaches are based on an extraction method, which identifies salient textual segments, most often sentences, in documents. Sentences containing the most salient concepts are selected, ordered and assembled according to their relevance to generate summaries (also called extracts).

Previous work on multi-document summarization includes, among others, centroid-based sentence selection [18], supervised learning [22], and information fusion [2]. The interested reader is directed to the DUC proceedings for more information on the various approaches. In this paper, we concentrate on graph-based ranking approaches. The rest of this section presents the previous work relevant to this type of summarization.

Approximately at the same time, Erkan and Radev [9] and Mihalcea [13] proposed to apply graph-based ranking algorithms to sentence extraction. The underlying idea is that of representing documents as graphs. Sentences are represented as nodes and relations between them, e.g. similarity measures, are represented as edges. Ranking algorithms are a way of deciding on the importance of a node, i.e. a sentence, based on the information drawn from the entire graph. Such approaches have several advantages. First, differently from most other methods, they do not require training data. Second, they are easily adaptable to other languages [14].

C. Cross-language Summarization

Cross-language summarization has received much attention recently and several approaches have been proposed. A natural way to go about this task would be to translate the documents prior to summarization, or to translate the generated summary. Orăsan and Chiorean [15] proposed to use the Maximal Marginal Relevance (MMR) method [6] to produce Romanian news summaries and then automatically translate them into English. More recently, Wan *et al.* [21] showed that incorporating translation quality scores in the summarization process increases both generated summary' content and readability. They focused on English-to-Chinese mono-document summarization and employed supervised learning to predict MT quality. In this study we will go a step further by incorporating MT confidence scores in cross-language multi-document summarization. Unlike the work of Wan *et al.*, our approach uses an unsupervised language-independent ranking algorithm for sentence selection [14].

[1]Document Understanding Conferences were conducted from 2000 to 2007 by the National Institute of Standards and Technology (NIST), `http://duc.nist.gov`

III. Method

In this section, we describe our method for cross-language multi-document summarization. We based our approach on a two-step summarization process which first scores each sentence, and then selects the top ranked sentences for inclusion in the summary. A preliminary step is added in order to translate each sentence and estimate the resulting translation quality. We modified the graph construction step to take advantage of the translation quality scores. Lastly, the French summary is constructed from the translation of the top ranked English sentences. Figure 1 presents an overview of the architecture of our proposed method.

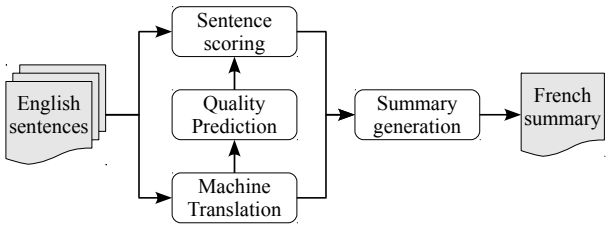

Fig. 1. Architecture of our proposed summarization system.

A. Pre-processing Documents and MT Quality Prediction

Each document in the cluster is segmented into sentences using the Punkt sentence boundary detection method [11] implemented in the NLTK toolkit [3]. All the English sentences were automatically translated into French using the Google translate service.[2]

An MT score is computed for each sentence to estimate both the translation accuracy and the fluency of the generated French sentences. This score aims at promoting in the summarization process sentences that can be easily read and understood by French speaking readers. In order to obtain it, we computed for each sentence 8 features that provide information on how difficult the source sentence is and how fluent the generated translation is:

- the source language sentence length in terms of words,
- the ratio of source and target lengths,
- the number of punctuation marks in the source language sentence,
- the proportion of the source numbers and punctuation symbols found in the target sentence,
- the perplexities of the source and the target sentences computed by 5-gram forward Language Models (LMs),
- the perplexities of the source and the target sentences computed by 2-gram backward LMs, i.e after reversing the word order of sentences.

[2]`http://translate.google.com`

These first four features belong to the most relevant features underlined by [20], among 84 features studied; the last four ones have already turned out to be effective for sentence-level confidence measures [19]. LMs are built using monolingual corpora of the news domain, made available for the WMT 10 workshop [5] and consisting of 991M English words and 325M French words. Perplexity scores are expected to reflect fluency, the use of 2-gram backward LMs addressing more specifically the detection of incorrect determinants or other function words. Contrary to other studies, we decided to focus on basic features that does not require any linguistic resources, such as parsers or dictionaries. Besides, features were restrained to scores computed only from the input sentence and its translated sentence, and therefore do not depend on the MT system used.

To predict MT quality from features, we adopt the ϵ-Support Vector Regression method (ϵ-SVR), already used for this purpose [21], [19]. In our experiments, we resort to the LIBSVM library [7] using the radial basis function as kernel, as recommended by the authors. The regression model depends on two parameters: an error cost c and a coefficient γ of the kernel function; their values have been optimized on a training corpus by grid search and cross-validation.

Ideally, the ϵ-SVR model should be trained on a corpus labeled with human judgments of MT output quality. Unfortunately, we are not aware of a large enough corpus of this kind for the English-French pair and producing MT judgments is a very slow process. We decided to resort instead to the automatic metric NIST [8] as an indicator of quality. Indeed, this metric have already been used in the past for this purpose [4], [20] and turned out to be more correlated with human judgments at the sentence level than other metrics such as the widely used BLEU [4]. Our training corpus was built from the reference translations provided in the news domain for the WMT workshops [5] from 2008 to 2010, which represents a set of 7,112 sentences. In order to assess the quality of the so-built model, we computed the Mean Squared Error (MSE) metric: $\frac{1}{N}\sum_{j=1}^{N}(y_j - \hat{y}_j)^2$, where N is the number of sentences, $\hat{y}$ is the prediction estimated by the regressor and y the actual value. On the 2,007 sentences made available for WMT 07 and kept for this purpose, we obtained a MSE of 0.456.

B. Sentence Scoring

We use a graph-based ranking approach to multi-document summarization. The first step is to construct a graph that represents the text. Let $G = (V, E)$ be a directed graph with the set of vertices (nodes) V and a set of directed edges E, where E is a subset of $V \times V$. Let $pred(V_i)$ be the set of vertices that point to the vertex V_i and $succ(V_i)$ the set of vertices that vertex V_i points to. A node is added to the graph for each sentence in the cluster. Connections (edges) between sentences (nodes) are defined in terms of similarity. We use the similarity measure proposed in [13], computed as a function of content overlap. The overlap of two sentences is the number of common tokens between the lexical representations of the two sentences, after stop words removal and stemming with the Porter stemmer. To avoid promoting long sentences, this number is normalized by the sentence lengths. Given $\mathrm{freq}(w, S)$ the frequency of word w in sentence S, the similarity between S_i and S_j is defined as:

$$Sim(S_i, S_j) = \frac{\sum_{w \in S_i, S_j} \mathrm{freq}(w, S_i) + \mathrm{freq}(w, S_j)}{\log(|S_i|) + \log(|S_j|)} \quad (1)$$

Graph-based ranking algorithms implements the concept of recommendation. Sentences are scored by taking into account global information recursively computed from the entire graph. In this study, we use an adaptation of the Google's PageRank ranking algorithm [16] to include edge weights:

$$p(V_i) = (1-d) + d \times \sum_{V_j \in pred(V_i)} \frac{Sim(S_i, S_j)}{\sum_{V_k \in succ(V_i)} Sim(S_k, S_i)} p(V_i) \quad (2)$$

where d is a "damping factor", which is typically chosen in the interval $[0.8, 0.9]$ (see [16]). This method, described in [13], is very similar to Lexical PageRank (LexRank) [9]. From a mathematical point of view, the PageRank algorithm computes the dominant eigenvector of the matrix representing the graph. We will use this method as baseline in our experiments.

C. Incorporating MT Quality Scores

In order to address the cross-language aspect, machine translation quality scores are introduced at the graph construction step. We modified Equation 1 to:

$$Sim_2(S_i, S_j) = Sim(S_i, S_j) \times Prediction(S_i) \quad (3)$$

where $Prediction(S_i)$ is the translation quality score of sentence S_i computed in Section III-A. Unlike the similarity measure defined by Equation 1 which is symmetric, this measure is directed. An accurate and fluent translated sentence would have its outgoing edge weights strengthen and hence would play a more central role in the graph. This way, sentences that are both informative and that are predicted to be accurately translated by the MT system will be selected.

We made some adaptations to the ranking algorithm to take advantage of the specificity of the documents. The position of a sentence within a document is a strong indicator of the importance of its content. This is especially true in newswire articles, which tend to always begin with a concise description of the subject of the article. Thus, double weight is given to all edges outgoing from a node corresponding to a leading sentence. Lastly, identical sentences (we keep only one occurrence) and sentences less than 5 word long are automatically dismissed.

D. Summary Generation

It is often the case that clusters of multiple documents, all related to the same topic, contain very similar or even identical sentences. To avoid such pairs of sentences, which may decrease both readability and content aspects of the summary, we have to use a redundancy removal method. Maximal Marginal Relevance (MMR) [6] is perhaps the most widely used redundancy removal technique. It consists in iteratively selecting summary sentences that are both informative and different from the already selected ones. In her work, Mihalcea introduces a maximum threshold on the sentence similarity measure [14]. Accordingly, at the graph construction step, no edge is added between nodes (sentences) whose similarity exceeds this threshold. In this study, we choose to use a two-step sentence selection method for maximizing the amount of information conveyed in the summary and minimizing the redundancy.

The second sentence selection step determines among the top scored sentences, as evaluated in the sentence ranking step, those which would make the best summary when combined together [10]. We first generate all the candidate summaries from combinations of the N sentences with the best relevance score that have the following properties: their combined number of characters does not exceed a threshold $\mathcal{T}$; no other sentences can be added while still remaining under a number of characters $\mathcal{T}$. Each candidate summary is then scored using a combination of word diversity (number of unique n-grams for $n \in [1, 2]$) and sentence relevance (sum of individual sentence scores). The sentences contained in the candidate summary with the best global score are the ones selected for the summary.

Summaries are constructed by sorting the selected sentences in chronological order to maximize temporal coherence. Sentences extracted from the oldest documents are displayed first. If two sentences are extracted from the same document, the original order within the document is kept.

IV. Results

In this section, we describe the details of our experimental protocol. We first give a description of the data set and the evaluation metrics we used. Then, we present the results obtained by our cross-language summarization system.

A. Experimental Settings

In this study, we used the document sets made available during the Document Understanding Conference (DUC) 2004 evaluation. DUC 2004 provided 50 English document clusters for generic multi-document summarization. Each cluster contains on average 10 newswire documents from the Associated Press and New York Times newswires. The task consists in generating short summaries representing all the content of the document set to some degree. Summaries must not exceeds 665 characters (alphanumerics, white spaces and punctuation included). This maximum length was derived from the manual summaries used in DUC 2003. We performed both automatic evaluation of content and manual evaluation of readability on a subset of the DUC 2004 data set made of 16 randomly selected clusters.

1) Automatic Evaluation: The majority of 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 generated summaries, we choose to use the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [12] evaluation toolkit, that has been found to be highly correlated with human judgments. ROUGE is a n-gram recall-based measure calculated as the number of overlapping n-grams between a candidate summary and a set of reference summaries. In our experiments, three metrics are computed: ROUGE-1 (unigram-based), ROUGE-2 (bigram-based) and ROUGE-SU4 (skip-bigram, allowing bigrams to be composed of non-contiguous words with as many as four words intervening). We run the version 1.5.5 of ROUGE with the default parameters[3] given by the DUC guidelines.

Reference English summaries for DUC 2004 were provided by NIST annotators. Four reference summaries were manually produced for each cluster. In our work, we focused on generating French summaries from English document sets. To be able to evaluate our method, we asked three annotators to translate the subset of 16 cluster's English reference summaries into French reference summaries. The translation instructions the annotators were given are fairly simple: each summary is to be translated sentence by sentence without introducing any kind of extraneous information (e.g. anaphora generation, proper name disambiguation or any sentence reduction technique). 64 reference summaries were translated this way, four for each cluster. The translators spent on average 15 minutes per summary (a total of more than 16 hours).

We have not restricted the size of the translated summaries to a given length. Accordingly, the length of the French reference summaries is on average 25% longer (in number of characters) than English ones. Similarly, our generation algorithm does not impose a maximum length on the French summaries but uses the total length of the corresponding English sentences. Lastly, we adapted the Porter stemmer embedded in the ROUGE evaluation package to correctly handle French words.

2) Manual Evaluation: The linguistic well-formedness of each summary is evaluated using a protocol similar to the one used during the DUC campaigns. We evaluate the readability aspect of the summaries on a five-point scale from 1 to 5, where 5 indicates that the summary is "easy to read", and 1 indicates that the summary is "hard to read". Annotators were asked to grade two randomly ordered summaries, one generated with the proposed method and the other obtained by translating the English output of a state-of-the-art approach

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

(described in Section III-B). Five annotators participated in the manual evaluation.

B. Monolingual Experiments

We first wanted to investigate the performance of the described method on a monolingual summarization task. Table I reports the automatic evaluation scores obtained on the DUC 2004 data set for different sentence scoring methods. *Graph-Sum* stands for the graph-based ranking method presented in Section III-B. Baseline results are obtained on summaries generated by taking the leading sentences of the most recent documents of the cluster, up to 665 characters (official baseline of DUC, identifier is 2). The table also lists the top performing system (DUC identifier is 65) at DUC 2004. We observe that the graph-based ranking approach achieves state-of-the-art performance, the difference with the best system is not statistically significant (paired Student's t-test of $\rho = 0.77$ for ROUGE-1, $\rho = 0.17$ for ROUGE-2 and $\rho = 0.57$ for ROUGE-SU4). By ways of comparison our system would have been ranked in the top 4 at the DUC 2004 campaign. Moreover, no post-processing was applied to the selected sentences leaving an important margin of progress.

TABLE I
ROUGE AVERAGE RECALL SCORES COMPUTED ON THE DUC 2004 DATA SET, THE RANK AMONG THE 35 PARTICIPANTS IS ALSO GIVEN. SCORES MARKED WITH † ARE STATISTICALLY SIGNIFICANT OVER THE BASELINE (PAIRED STUDENT'S T-TEST WITH $\rho < 0.001$)

System	ROUGE-1	**rank**	ROUGE-2	**rank**	ROUGE-SU4	**rank**
1^{st} system	0.38244†	1	0.09218†	1	0.13323†	1
Graph-Sum	0.38052†	2	0.08566†	4	0.13114†	3
Baseline	0.32381	26	0.06406	25	0.10291	29

C. Cross-language Experiments

In this second series of experiments, we evaluated our method for cross-language multi-document summarization. Baseline results are obtained by translating the English output of the graph-based ranking approach (described in Section III-B). The automatic ROUGE evaluation scores are presented in Table II. We observe a small improvement in ROUGE-2 and ROUGE-SU4 for our method. Nevertheless, this increase is not significant. This result can be explained by the fact that MT quality scores can promote inside the summary some sentences that are less informative but more understandable and readable.

TABLE II
ROUGE AVERAGE RECALL SCORES COMPUTED ON THE FRENCH TRANSLATED SUBSET OF THE DUC 2004 DATA SET

System	ROUGE-1	ROUGE-2	ROUGE-SU4
Baseline	0.39704	0.10249	0.13711
Our method	0.39624	0.10687	0.13877

We then evaluated the linguistic well-formedness of the summaries generated with our proposed method. Table III shows the manual evaluation results on the subset of 16 clusters. The average score given by each human judge is also given. We observe that the proposed approach obtains better readability scores. All annotators agree that our method produces more easy-to-read summaries than the baseline. This result indicates that MT quality scores are useful for selecting more readable sentences. An example of generated summaries is given in Appendix 1. Overall, results show that our method can enhance the readability of the generated summaries without degrading their informativity. However, the average readability scores are relatively low. An analysis of the errors observed in French summaries leads us to think that pre-processing source sentences (e.g. removing ungrammatical sentences) can be a first step to filter out erroneous sentences.

TABLE III
READABILITY SCORES OF OUR PROPOSED METHOD COMPARED TO THE STANDARD GRAPH-BASE RANKING APPROACH (BASELINE). SCORES ARE ON A FIVE-POINT SCALE FROM 1 TO 5, WHERE 5 INDICATES THAT THE SUMMARY IS "EASY TO READ", AND 1 IS "HARD TO READ"

Annotator	**Readability**	
	Baseline	Our method
Annotator 1	2.44	2.50
Annotator 2	1.56	1.63
Annotator 3	1.75	2.31
Annotator 4	3.06	3.31
Annotator 5	1.50	1.63
Average	2.06	2.28

V. CONCLUSION AND FUTURE WORK

In this paper, we presented a graph-based approach to cross-language multi-document summarization. We proposed to introduce machine translation quality scores at the graph construction step. Automatically translated sentences that are both fluent and informative are then selected by our ranking algorithm. We evaluated our approach on a manually translated subset of 16 clusters from the DUC 2004 data set. Results show that our approach enhances the readability of the generated summaries without degrading their content.

In future work, we intend to expand the set of reference summaries by translating the entire DUC 2004 data set. We also plan to extend the evaluation to other languages. The manually translated French summaries introduced in this paper, along with the manual given to the group of translators, is available for download on request.

REFERENCES

[1] C. Banea, A. Moschitti, S. Somasundaran, and F. M. Zanzotto, Eds., *Proceedings of TextGraphs-5 Workshop*, Uppsala University. Uppsala, Sweden: ACL, 2010. [Online]. Available: http://www.aclweb.org/anthology/W10-23

[2] J. Blatz, E. Fitzgerald, G. Foster, S. Gandrabur, C. Goutte, A. Kulesza, A. Sanchis, and N. Ueffing, "Confidence estimation for machine translation," Johns Hopkins University, Batimore, MD, USA, Tech. Rep., 2003.

[3] S. Raybaud, D. Langlois, and K. Smaïli, "Efficient combination of confidence measures for machine translation," in *Proceedings of Interspeech 2009 conference*, Brighton, UK, 2009, pp. 424–427.
[4] L. Specia, N. Cancedda, M. Dymetman, M. Turchi, and N. Cristianini, "Estimating the sentence-level quality of machine translation systems," in *Proceedings of EAMT 2009 conference*, Barcelona, Spain, 2009, pp. 28–35.
[5] C. B. Quirk, "Training a sentence-level machine translation confidence measure," in *Proceedings of LREC 2004 conference*, Lisbon, Portugal, 2004, pp. 825–828.
[6] D. Radev, H. Jing, M. Sty, and D. Tam, "Centroid-based summarization of multiple documents," *Information Processing & Management*, vol. 40, no. 6, pp. 919–938, 2004.
[7] K.-F. Wong, M. Wu, and W. Li, "Extractive summarization using supervised and semi-supervised learning," in *Proceedings of Coling 2008 conference*, Manchester, UK, 2008, pp. 985–992. [Online]. Available: http://www.aclweb.org/anthology/C08-1124
[8] R. Barzilay, K. R. McKeown, and M. Elhadad, "Information fusion in the context of multi-document summarization," in *Proceedings of ACL 1999 conference*, College Park, MD, USA, 1999, pp. 550–557. [Online]. Available: http://www.aclweb.org/anthology/P99-1071
[9] G. Erkan and D. Radev, "LexRank: Graph-based lexical centrality as salience in text summarization," *JAIR*, vol. 22, no. 1, pp. 457–479, 2004.
[10] R. Mihalcea, "Graph-based ranking algorithms for sentence extraction, applied to text summarization," in *Proceedings of ACL 2004 conference*, Barcelona, Spain, July 2004, pp. 170–173.
[11] R. Mihalcea and P. Tarau, "A language independent algorithm for single and multiple document summarization," in *Proceedings of IJCNLP 2005 conference*, vol. 5, Jeju Island, South Korea, 2005.
[12] C. Orăsan and O. A. Chiorean, "Evaluation of a cross-lingual romanian-english multi-document summariser," in *Proceedings of LREC 2008 conference*, Marrakech, Morocco, 2008. [Online]. Available: http://clg.wlv.ac.uk/papers/539_paper.pdf
[13] J. Carbonell and J. Goldstein, "The use of MMR, diversity-based reranking for reordering documents and producing summaries," in *Proceedings of SIGIR 1998 conference*. ACM, 1998, pp. 335–336.
[14] X. Wan, H. Li, and J. Xiao, "Cross-language document summarization based on machine translation quality prediction," in *Proceedings of ACL 2010 conference*, Uppsala, Sweden, 2010, pp. 917–926. [Online]. Available: http://www.aclweb.org/anthology/P10-1094
[15] T. Kiss and J. Strunk, "Unsupervised multilingual sentence boundary detection," *Computational Linguistics*, vol. 32, no. 4, pp. 485–525, 2006.
[16] S. Bird and E. Loper, "Nltk: The natural language toolkit," in *Proceedings of ACL 2004 conference*, Barcelona, Spain, 2004, pp. 214–217.
[17] C. Callison-Burch, P. Koehn, C. Monz, K. Peterson, M. Przybocki, and O. Zaidan, "Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation," in *Proceedings of the Joint 5th Workshop on Statistical Machine Translation and MetricsMATR (WMT)*, Uppsala, Sweden, 2010, pp. 17–53. [Online]. Available: http://www.aclweb.org/anthology/W10-1703
[18] C.-C. Chang and C.-J. Lin, *LIBSVM: a library for support vector machines*, 2001, software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.
[19] G. Doddington, "Automatic evaluation of machine translation quality using n-gram co-occurrence statistics," in *Proceedings of HLT 2002 conference*, San Diego, CA, USA, 2002, pp. 138–145.
[20] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web," Stanford Digital Library Technologies Project, Tech. Rep., 1998.
[21] P. Genest, G. Lapalme, L. Nerima, and E. Wehrli, "A symbolic summarizer with 2 steps of sentence selection for tac 2009," in *Proceedings of TAC 2009 Workshop*, Gaithersburg, MD, USA, 2009.
[22] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text Summarization Branches Out: Proceedings of ACL-04 Workshop*, S. S. Marie-Francine Moens, Ed., Barcelona, Spain, 2004, pp. 74–81.

APPENDIX 1

TABLE IV
EXAMPLE OF FRENCH SUMMARIES GENERATED FOR THE DUC CLUSTER D30007T BY THE BASELINE AND THE PROPOSED APPROACH

Baseline (average readability score of 2.4)

Après une journée de combats, les rebelles congolais a annoncé dimanche avoir conclu Kindu, la ville stratégique et à la base dans l'est du Congo utilisé par le gouvernement pour mettre fin à leurs avances. (*After a day of fighting, Congolese rebels said Sunday they had entered Kindu, the strategic town and airbase in eastern Congo used by the government to halt their advances.*) Etienne Ngangura, un porte-parole des rebelles, a déclaré les combattants rebelles se trouvaient dans Kindu et avait pris le côté, grande base aérienne, 380 km (235 miles) à l'ouest de Goma, le fief des rebelles. (*Etienne Ngangura, a rebel spokesman, said the rebel fighters were inside Kindu and had taken the adjacent, large airbase, 380 kilometers (235 miles) west of Goma, the rebel stronghold.*) "Nos soldats sont dans la ville et les combats se poursuivent", le commandant de bataillon rebelle Arthur Mulunda a déclaré à Kalima, à 80 kilomètres (50 miles) au nord de Kindu. (*"Our soldiers are in the town and the fighting is continuing" rebel battalion commander Arthur Mulunda said in Kalima, 80 kilometers (50 miles) northeast of Kindu*) Le samedi, les rebelles ont dit qu'ils ont abattu un Boeing 727 Congolais qui tentait d'atterrir à la base aérienne de Kindu avec 40 troupes et de munitions. (*On Saturday, the rebels said they shot down a Congolese Boeing 727 which was attempting to land at Kindu air base with 40 troops and ammunition*)

Our method (average readability score of 3.2)

Les rebelles ont attaqué un village dans l'ouest de l'Ouganda et a tué six civils devant des soldats contraints de rebrousser chemin, une porte-parole militaire a déclaré jeudi. (*Rebels attacked a village in western Uganda and killed six civilians before soldiers drove them off, a military spokesman said Thursday*) Etienne Ngangura, un porte-parole des rebelles, a déclaré les combattants rebelles se trouvaient dans Kindu et avait pris le côté, grande base aérienne, 380 km (235 miles) à l'ouest de Goma, le fief des rebelles. (*Etienne Ngangura, a rebel spokesman, said the rebel fighters were inside Kindu and had taken the adjacent, large airbase, 380 kilometers (235 miles) west of Goma, the rebel stronghold*) Les commandants rebelles, a déclaré mardi qu'ils étaient sur le point d'envahir une importante base aérienne détenue par le gouvernement au Congo Est, une bataille qui pourrait déterminer le futur de la guerre de deux mois congolais. (*Rebel commanders said Tuesday they were poised to overrun an important government-held air base in eastern Congo, a battle that could determine the future of the two-month Congolese war*) Les rebelles dans l'est du Congo a déclaré samedi qu'ils ont abattu un avion de ligne transportant 40 soldats du gouvernement dans un aéroport stratégique face à un assaut des rebelles. (*Rebels in eastern Congo on Saturday said they shot down a passenger jet ferrying 40 government soldiers into a strategic airport facing a rebel assault*)