Text Summarization by Sentence Extraction and Syntactic Pruning

Michel Gagnon
Département de génie informatique
École polytechnique de Montréal
Michel.Gagnon@polymtl.ca

Lyne Da Sylva
École de bibliothéconomie et des sciences de l'information
Université de Montréal
Lyne.Da.Sylva@UMontreal.CA

Abstract

We present a hybrid method for text summarization, combining sentence extraction and syntactic pruning of extracted sentences. The syntactic pruning is done based on a complete dependency-grammar analysis of sentences, performed by the grammar developed within a commercial French grammar checking software package, le Correcteur 101. Subtrees in the syntactic analysis are pruned when they are labelled with targeted relations. Evaluation is performed on a corpus of various texts. The reduction ratio of extracted sentences averages around 74%, while retaining grammaticality or readability in a proportion of over 64%. Given these first results on a limited set of syntactic relations, this shows promise for a text summarization method.

1 Introduction

This paper deals with text summarization, whose goal is to produce a shorter version of a source text, while still retaining its main semantic content. Research in this field is flourishing (see namely Mani, 2001; Minel, 2004; NIST, 2005); it is motivated by the increasing size and availability of digital documents, and the necessity for more efficient methods of information retrieval and assimilation.

Methods of automatic summarization include extracting (summarizing by using a limited number of sentenced extracted from the original text) and abstracting (producing a new, shorter text). Extraction algorithms have a strong tendency to select long sentences from the text (since word frequency and distribution are often crucial, and are higher in long sentences even when sentence length is factored in). Shortening the extracted sentences can be a way to further reduce the resulting summary, provided that the (essential) meaning of the sentence is preserved. Such summaries can presumably allow for shorter reading time. We have developed a hybrid method which combines sentence extraction and reduction of these sentences.

After presenting our objectives and previous related work, this article details the methodology, and then presents and discusses experimental results. The conclusion outlines future work.

2 Objectives

Three objectives are sought in this paper. First, we present the method for text reduction based on a hybrid approach which combines sentence extraction and syntactic pruning of extracted sentences. We describe each component and its contribution to the system as a whole.

Secondly, although we recognize the existence of numerous resources for the analysis and summarization of English texts (and the evaluation thereof), equivalent systems for French are scarce. Given resources at our disposal, namely a broad-coverage grammar for French, we developed a system for summarization of French texts.

Finally, we present an evaluation of the hybrid approach on a collection of texts; this aims to determine whether, with a greater rate of compression, the resulting reduced sentences preserve the essential semantics of the original sentences. Success would suggest this approach has potential as a summarization method.

3 Related work

3.1 Extracting

Text extracts are produced by identifying "interesting" sentences in the source document and simply joining them to produce what is hoped to be a legible summary. Various methods exist to determine which sentences should be extracted and a number of commercial systems using these are available (Copernic Summarizer, Microsoft Word's summarizer, Pertinence Sumarizer, Xerox InXight, ...).

Our implementation, Cortex, was developed at the École Polytechnique de Montréal. It compares favourably with other extractors (see Torres-Moreno et al., 2004). Like most other systems, it is based on a matrix of word frequency in the text; but it uses an original algorithm to combine various statistical measures.

Extracting is a simple, robust method, but it suffers from a number of problems; the one we focus on is the fact that extracted sentences may be wordy and not quite reach the goal of summarizing the document sufficiently.

3.2 Abstracting

An abstract is "a summary at least some of whose material is not present in the input" (Mani, 2001:129). An abstract may start by reducing sentences from the source text, joining sentence fragments, generalizing, etc. This method has greater potential for increasingly readable summaries. Although the most ambitious abstracting methods require a full analysis of the input text, much previous work has relied on limited analysis, for instance information extraction templates (Rau et al., 1989; Paice and Jones, 1993; McKeown and Radev, 1995), rhetorical structure trees (Marcu, 1996, 1999) and a comparison of a noisy-channel and a decision-tree approach (Knight and Marcu, 2002). Some researchers have tried to identify linguistic text reduction techniques which preserve meaning (Jing and McKeown, 1999; Saggion and Lapalme, 2000). These techniques vary considerably and some are much harder to implement than others; however, all require a fairly good syntactic analysis of the source text. This implies having a wide-coverage grammar, a robust parser, and generation techniques which defy most existing systems.

3.3 Text reduction based on syntactic analysis

There is indeed a potential for useful text reduction given a robust syntactic analysis. Limited work has been performed in this area. Grefenstette (1998) experiments with sentence reduction based on a syntactic analysis provided by a robust parser (Grefenstette, 1996). He defines various levels of "informativeness": proper names are considered most informative, then common nouns, then adjectives, determiners and subordinate clauses, etc. On this "informativeness hierarchy", text compaction levels are defined, where level 1 only keeps proper nouns (which we would consider indexing via named-entities, and not summarization), level 2 keeps subjects and objects, level 3 keeps main verbs, and level 4 keeps prepositional phrases but no subordinate clause, etc. Level 4 is the first one where one can expect to have grammatical sentences. These results can be improved given a more sophisticated syntactic pruning methodology.

Mani et al. (1999) compress extracted sentences based on a phrase-structure syntax analysis indirectly based on Penn Treebank data; pruning is performed (among other operations) on certain types of phrases in specific configurations, including parentheticals, sentence-initial PPs and adverbial phrases such as "In particular,", "Accordingly," "In conclusion," etc.

We are interested in exploring the potential of pruning a dependency-syntax analysis, which is based on a representation which directly encodes grammatical relations and not merely phrase structure. We believe the latter allows a better characterization of the sub-parts of the tree that can safely be pruned while retaining essential meaning. Indeed, grammatical relations such as subject and direct object should correlate with central parts of the sentence, whereas subordinate clauses and temporal or locative adjuncts should correlate with peripheral information. Pruning decisions based on this type of criteria seem better motivated than those based on phrase structure. (Note that this is still different from pruning a semantic representation (e.g.Fiszman et al., 2004)).

Moreover, we are aware of no work on French dealing with sentence reduction based on a dependency analysis. We currently do have access to the source code for a robust, wide-coverage grammar of French, developed within a commercial grammar-checking product (Le Correcteur 101™, by Machina Sapiens and now Lingua Technologies[1]). The grammar is dependency-based: syntactic trees consist of nodes corresponding to the words of the sentence, and links between nodes are

[1] www.LinguaTechnologies.com

labelled with grammatical relations (of the type “subject”, “direct object”, “subordinate clause”, “noun complement”, etc.).

The grammar aims to perform a complete syntactic analysis of the sentence (see Figure 1 for an indented presentation). In case of failure (due to severe writer error or to limits of the grammar), it provides a series of partial analyses of fragments of the sentence. In all cases, Correcteur 101 ranks analyses using an in-house weighting mechanism.

```
Les médias sont-ils responsables de
l’efficacité des publicités qu’ils véhi-
culent ?

Arbre  sont/verbe
  Sujet  les médias/nom
  RepriseSujet  ils/pronPers
  Attrib  responsables/adj
   ComplAdj de l'efficacité/nom
     ComplNom  des publicités/nom
      Relat  véhiculent/verbe
       ObjetDirect  qu'/pronRelat
       Sujet  ils/pronPers
  FinProp  ?/ponctFinale
```

Figure 1. Sample dependency tree: main verb is labelled as “Arbre”. Some sub-trees are simplified.

This grammar bears many important advantages. In addition to its large coverage, it is able to provide a full analysis even with erroneous input. Its 80,000 lines of C++ code represent many person-years of development; the grammar consists of over 2500 grammar rules and a dictionary containing over 88,000 entries.

The detailed analysis produced by the grammar can be the basis of syntactic pruning for text reduction (this is illustrated in Figure 2).

Because of its use in grammar (and spelling) correction, the grammar is highly robust. It does, however, have peculiarities which we discuss below. In brief, certain linguistic phenomena are ignored when they have no effect on correction. Note that other recent work (Coulombe, Doll and Drouin, 2005) also uses this grammar in a non-correcting context, pertaining to controlled languages.

```
[Dans le monde en pleine effervescence
d'Internet, ]locAdj l'arrivée de HotWired
marque le début de la cybermédiatisation
[, le premier véritable média sur Inter-
net]app.
➔
L'arrivée de HotWired marque le début de
la cybermédiatisation.
```

Figure 2. Sample reduction: locative adjunct (locAdj) and apposition (app).

4 Methodology

We have developed a prototype which first applies a sentence extraction algorithm (Cortex) to spot the most prominent sentences of a source text, then performs sentence reduction using syntactic pruning of the automatically extracted sentences.

A variation on this approach is to use syntactic pruning to improve initial sentence selection (Siddharthan et al., 2004), which we have also experimented with (but which we do not report on in this paper).

The version of Cortex used here combines four metrics: first, word similarity of each sentence with the title; second, the position of each sentence within the text; and the last two metrics evaluate (in different ways) the interaction among sentences in the text, by considering shared words. Summary size is set to 10%, in terms of sentences.

Our method of syntactic pruning improves on Grefenstette’s experiments. His did not involve sentence extraction, only reduction. Also, his definition of compaction levels uses a analysis which is less fine-grained than what is possible with the dependency grammar of 101. And we should be able to maintain the sentence’s grammaticality, insofar as we prune only subordinate material, and never the main verb of the sentence.

For the syntactic pruning, we adapted the output of Le Correcteur 101 to produce parses corresponding to the full tree and to the pruned tree.

The grammar of 101 is used in its entirety. Extracted sentences are submitted one by one and a complete syntactic analysis of each is performed. Although 101 usually supplies all plausible analyses (more than one, in the case of ambiguous syntactic structures), for our prototype we use only the top-ranking one. This has some limitations: sometimes two or more analyses share the same rank, or the highest-ranking one is not the correct one (as determined by a human judge). Our prototype systematically chooses the first one, regardless. The impact of incorrect analyses is great, as it radically changes results: complements may be related by a different relation, and thus the reduction performed is not the one intended.

Then a filtering operation follows, which removes a sub-tree in the dependency graph when the labelled relation corresponds to one in a predefined list. The entire sub-tree is removed, thus efficiently applying the reduction operation. An external file contains the syntactic relations that trigger reduction, which allows for easy testing of various sets of relations. A preliminary test was performed using a wide number of relations. Only obligatory complements and phrasal specifiers (such as determiners) were kept. This resulted in a large reduction, producing much shorter sentences which however tended to be ungrammatical. It was determined that a much more focused approach would have a better chance of reducing the text while still preserving important elements and grammaticality.

For the final run, only the following relations were pruned: prepositional complements of the verb, subordinate clauses, noun appositions and interpolated clauses ("incises", in French). This is encoded with 6 relations, out of the 246 relations used by 101. However, 101 attaches all prepositional complements of the verb with the same relation as optional adjuncts such as time, place, etc. This was done during the development of the grammar to reduce the number of possible analyses of sentences with prepositional phrases (given that an ambiguity of this type is never relevant for correction purposes). To circumvent this problem, a provision was made in our prototype for obligatory complements of the verb (for example, "à Montréal" in "Il habite à Montréal"). The latter must not be pruned, to avoid incomplete and ungrammatical verb phrases. Since this is not encoded in the lexical entries used by 101, it had to be added; for our tests, we hand-coded only a number of such prepositional complements, for the verbs identified in our corpus. We call these "anti-filters", as their purpose is to prevent the filtering of the corresponding complement.

The test corpus consisted of 10 texts of various sizes and genres (see Table 1). The methodology used explains the small size of the corpus: evaluation necessitated a careful, manual examination of all extracted sentences (original and pruned). Some of the texts are quite long ("opusdei", for example). No evaluation corpus was at our disposal for this collection of dependency analyses of French texts and their summaries. As to the variety of the corpus in terms of genres (which may hinder the detection of systematic errors), quite simply this was the corpus that had been used in a previous evaluation of Cortex and we could use its selection of extracted sentences. We also believe that using a single genre, although it may help to detect trends within the genre, can certainly also introduce bias (as for example using a journalistic corpus entails).

Ident.	# sentences	# words	Genre
cybermedia	62	1276	Journalistic
Jaccuse	207	4912	Political pamphlet
Durham	210	6515	Report
Lavie	139	4373	Popular science
Epicier	191	3438	Literary
Epistemo	230	5506	Monograph
Football	102	2761	Essay
Science	225	5627	Popular science
Opusdei	443	8529	Literary
Travail	244	8264	Monograph

Table 1. Details of the test corpus.

5 Results

Each sentence was examined to determine if (i) it had been pruned (ii) whether the result was "good" and (iii) whether it was grammatical. Good reductions are those which are either perfect (i.e. the main semantic content of the original sentence is retained in the reduction – see Figure 2) or acceptable (i.e. a part of the semantic content is lost, but the meaning of the reduced sentence is compatible with that of the original). Bad reductions were those where crucial semantic content was lost. Below is one example of the last two types.

```
Acceptable:
Le Soleil lui-même a de nombreuses répliques dans le ciel ;
→
Le Soleil lui-même a de nombreuses répliques ;
```

```
Bad:
Je n'entretiens aucun doute sur le caractère national qui doit être donné au Bas-Canada ;
→
Je n'entretiens aucun doute ;
```

Some cases were ungrammatical (see example below); this happened when our system removed elements which were in fact obligatory, but labelled with one of the relations subject to pruning (this may have been because 101's analysis was wrong).

```
les objets ont pour fonction de
stabiliser la vie humaine.

→

les objets ont de stabiliser la
vie humaine.
```

At other times, it was some artefact of 101's application of corrections or reductions. These cases were often ungrammatical but still considered acceptable, as in the example below.

```
Je désire que la vérité éclate
et que si vraiment, comme tout
semble le faire croire, c'est
cet épicier qui était le diable,
il est convenablement châtié.

→

Je désire que la vérité éclate
et que si vraiment, c'est cet
épicier qui était le diable, il
est convenablement châtié.
```

5.1 Statistics

We calculated the reduction rate for each sentence (i.e. the size of the pruned sentence, in words, compared to the number of words of the original sentence), then computed the figure globally for each text. We examined each reduced sentence and produced our own, "ideal" reduction (this is a subjective evaluation, but by a trained computational linguist); we calculated thus the ideal reduction rate.

Table 2 presents the reduction rate obtained from our experiments. The first column shows the rate obtained with our reduction module, whereas the second column contains the values we should expect according to our human evaluation. To obtain the statistics given in the third column, we restrict the computation of the reduction rate to only those sentences that have been correctly reduced. Similarly, we give in the fourth column the reduction rate for the sentences that have been erroneously reduced.

Table 3 shows, in the first column, the ratio of reduced sentences (unreduced sentences either contained no relations subject to pruning, or were protected by anti-filters), in comparison with the total number of sentences in the original summary. The second column shows the proportion of correctly reduced sentences, among the sentences that have been pruned.

	Reduction rate (in terms of words) of extracted sentences (%)			
Identifier	Obtained	Ideal	Among good reductions	Among "bad" reductions
cybermedia	69	71	70	34
Jaccuse	68	67	65	38
Durham	68	71	61	16
Lavie	81	70	77	52
Epicier	74	66	65	52
Epistemo	73	81	70	47
Football	80	80	65	60
Science	74	70	65	62
Opusdei	82	86	63	68
Travail	73	76	63	59
Average	74	74	66	49

Table 2: Reduction rate

We make the following observations. First, considering the average reduction rate, we see that 25% of words have been removed. This shows great potential for useful reduction of an extract. Note that this result correspond to the "ideal" reduction rate, but looking at the figures of Table 3, we see that about two thirds of the sentences are incorrectly reduced. This shows that correct pruning in some sentences is offset by incorrect pruning in other sentences.

Secondly, "good" reductions are those where the reduction is smaller (66% of words have been retained); for bad reductions, only 49% of words have been retained. Thus, reduction rate is higher when done incorrectly. One consequence of this is that avoiding bad reduction may cause some degradation in the overall reduction rate. But, since the ideal reduction rate is equal to the one obtained by our system, and we still have many sentences that could be reduced (we found that 41% of non-reduced sentences could be reduced according to human judgment), we may expect that this degradation would be compensated for by refining the reduction process.

Identifier	Number of reduced sentences / number of sentences (%)	Number of correctly reduced sentences / number of reduced sentences (%)
cybermedia	71	60
Jaccuse	86	83
Durham	82	89
Lavie	79	64
Epicier	65	62
Epistemo	70	50
Football	73	63
Science	65	47
Opusdei	40	72
Travail	72	56
Average	70	64

Table 3: Ratio of reduced sentences

Thirdly, as shown in Table 3, for 6 out of 10 texts, the proportion of reductions which are deemed "good" (perfect or acceptable) is over 60%.

In addition, we examined the bad reductions. Once again, a high proportion of these sentences, (57%) could have been reduced, but differently. This suggests that our choice of targeted relations was neither overly wide nor too narrow, but that finer tuning is necessary.

The compression rate for the summary produced by Cortex has been fixed to a 10% value, but since it is calculated in terms of number of sentences, the real compression rate is in fact 17,5% on average. By coupling Cortex with sentence reduction, the compression rate drops to 12,6%, which is closer to the desired value. Thus, syntactic pruning somehow compensates for the inaccuracy of compression rate based on the number of sentences and circumvents Cortex's proneness to select long sentences.

Finally, of the 37 sentences deemed ungrammatical after processing, 14 (38%) were still deemed good reductions (some of the ungrammaticalities were due to quirks in 101's behaviour).

5.2 Some problems with the grammar

As was to be expected, a number of incorrectly reduced sentences are due to the fact that the correct analysis was not the top-ranking one (although it was quite often provided by 101, as another alternative). And when the grammar had trouble finding the right analysis, it sometimes suggested corrections that were inappropriate. Finally, in 16% of cases, 101 was unable to give a complete analysis, but provided analyses of fragments instead.

A problem occurs with sentences containing specific types of complements clauses. Verbs which require an if-clause ("completive en si"), such as "se demander" ("to wonder"), have their complement labelled with a subordinate clause relation (again, to reduce the number of unnecessary ambiguous analyses in 101). This clause is an obligatory complement and should not be pruned (just as direct objects and predicative adjectives are not), but is pruned due to the "subordinate clause" label it receives. This would require a more complex pattern recognition, since two relations are involved (see Figure 4), which is not allowed in our anti-filter.

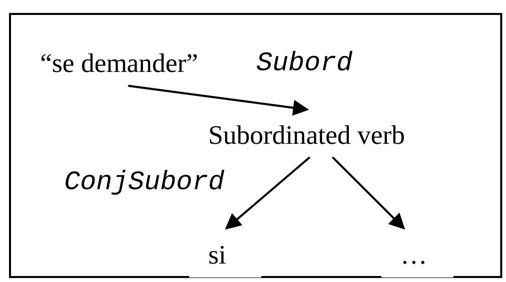

Figure 4. Pattern for if-clauses

Finally, our sub-tree pruning method would benefit from more use of context. In addition to the anti-filters for verb complements, it seems necessary at times to use contextual elements in addition to the relation name, such as properties of the father or of the sons involved in the relation.

6 Discussion

By a closer inspection of sentences that are incorrectly reduced, we found that in 37% of the cases, a good reduction would necessitate major changes in our model, or some semantic information that is beyond the scope of the parser (see Figure 5).

```
 L'histoire des sciences est venue
quant à elle montrer que le vrai et
le faux ne peuvent être pensés dans
une opposition rigide et unilatéra-
le.
→
 L'histoire des sciences est venue
montrer que le vrai et le faux ne
peuvent être pensés.
```

Figure 5: Sample unrecoverable reductions

For the remaining sentences, small improvements in the reduction process are required. In some cases, we would only have to add some entries in the anti-filter. Another very frequent situation is the pruning of a subordinated clause that should be prevented. A typical example is the *if*-clause discussed in last section, which occurs very frequently. Even when it is not part of the verb arguments, it should never be pruned considering its crucial role in the semantic interpretation. It is not possible to protect this kind of subordinated clause with the existing anti-filter, because we need a more complex pattern to recognize this case, as illustrated in Figure 4. Like subordination attachment, most of the other cases of bad reduction could be avoided by using slightly more complex patterns which would encompass more than a single relation between two words.

Since sentences judged to be incorrectly reduced are those which undergo more reduction, it suggests that a threshold could be determined (which could take the relation name into account). Also, we could use the fact that 101 can detect errors in sentences to submit reduced sentences to 101, and refuse reduction when the result is ungrammatical (i.e. correct by 101 itself).

7 Conclusion

We have proposed a method for text summarization which combines extraction and abstraction. The reduction rate achieved by our system (about 74% reduction of sentences extracted) shows great promise for a hybrid summarization approach.

Future work will follow a number of directions. First, we will examine which other relations can safely be pruned. A good candidate seems to be the attributive relation for adjectives ("adjectifs épithètes"), as many are usually not necessary. However, this requires a more careful study, as was apparent in our preliminary test). On the contrary, parts of idioms (ex. "prendre le taureau par les cornes") should be recognized as such and never be pruned. This is possible, given 101's analysis of idioms, but was not taken into account for these experiments.

Also, it may be that some "light" verbs (such as "avoir", "être", "faire") should never be separated from their complements; this requires further study (indeed, "être" is so frequent that to avoid reduction in its presence may lead to very little pruning).

We plan to add entries to our anti-filter. But we have yet to see how our reduction rate will be affected by adding a large number of anti-filter specifications. Since we are not using any semantic information, the anti-filter must be permissive, in the sense that if a verb has many senses with different sub-categorizations, the corresponding entries will pertain to the same lexical entry of the verb. With a large-scale anti-filter, this may affect the reduction rate.

The pattern matching rule used to target subtrees to be pruned is currently limited to a local tree with a single son; other pattern matching could be explored. But we must be wary of the trade-off between gained expressivity in the pattern matching process and computational performance.

Finally, we also have access to a similar dependency grammar for English (developed as part of a grammar checker as well). Its coverage is not as wide as that of Le Correcteur 101, but it has the advantage of having its lexicon of verbs completely specified as to obligatory prepositional complements. For this reason, we intend to pursue experiments on English texts.

Our working hypothesis for this work was that pruning of a dependency structure can be helpful in reducing sentences for summarization, while retaining essential meaning. It will be interesting to test this on output from other parsers using input with which it may be easier to perform an evaluation, using existing human produced summaries.

Acknowledgements

This researched was funded by a grant from the National Science and Engineering Research Council awarded to Lyne Da Sylva. Many thanks to Claude Coulombe, president of Lingua Technologies, for use of the source code for Le Correcteur 101. We also wish to thank our research assistant, Frédéric Doll.

References

Coulombe, Claude; Doll, Frédéric; Drouin, Patrick. Intégration d'un analyseur syntaxique à large couverture dans un outil de langage contrôlé en français. In : *Lingvisticae Investigationes* (Machine Translation, Controlled Languages and Specialised Languages), 28(1) (in press).

Fiszman, Marcelo; Rindflesch, Thomas C.; Kilicoglu, Halil. Abstraction Summarization for Managing the Biomedical Research Literature. In: *Proceedings of the Computational Lexical Semantics Workshop*, HLT-NAACL 2004, pp. 76-83, May 2-7, 2004, Boston, Massachusetts, USA.

Grefenstette, Gregory. Producing Intelligent Telegraphic Text Reduction to provide an Audio Scanning Service for the Blind. In : *Working Notes of the Workshop on Intelligent Text Summarization*, Menlo Park, CA: American Association for Artificial Intelligence Spring Symposium Series, 1998, pp. 111-117.

Grefenstette, Gregory. Light Parsing as Finite-State Filtering. In : *Proceedings ECAI '96 Workshop on Extended Finite State Models of Language*, August 11-12 1996, Budapest.

Jing, H. ; McKeown, K. The decomposition of human-written summary sentences. In : *Proceedings of the 22nd International Conference on Research and Development in Information Retrieval (SIGIR'99)*, New York: Association for Computing Machinery, 1999, pp. 129-136.

Knight, Kevin; Marcu, Daniel. Summarization beyond sentence extraction: A probabilistic approach to sentence compression. *Artificial Intelligence*, 139, 2002, pp. 91-107.

Mani, Inderjeet; Gates, Barbara; Bloedorn, Eric. Improving Summaries by Revising Them. In: *Proceedings of the 37th Annual Meeting of the ACL*, pp. 558-565, University of Maryland, College Park, Maryland, USA, 20 – 26 June 1999.

Mani, Inderjeet. *Automatic summarization*. Amsterdam; Philadelphia: John Benjamins, 2001.

Mani, Inderjeet ; Maybury, Mark T. *Advances in Automatic Text Summarization*. Cambridge, Mass.: MIT Press. 1999.

Marcu, Daniel. Discourse Trees are Good Indicators of Importance in Text. In: Mani, Inderjeet ; Mark Maybury. *Advances in Automatic Text Summarization*, Cambridge, Mass.: MIT Press, 1999, pp. 123-136.

Marcu, Daniel. Building up rhetorical structure trees. In : *Proceedings of the Thirteenth National Conference on Artificial Intelligence*, 1996, pp. 1069-1074.

McKeown, K.; Dragomir Radev. Generating Summaries of Multiple News Articles. In: *Proceedings of the 18th International Conference on Research and Development in Information Retrieval (SIGIR'95)*, Seattle, Wa.: Association for Computing Machinery, 1995, pp. 74-82.

Minel, Jean-Luc. Résumé automatique de textes. *Traitement automatique des langues*, 45(1), 2004.

NIST. *Document Understanding Conferences – Introduction.* [http://www-nlpir.nist.gov/projects /duc/intro. html]

Pinto, Maria; Galvez, Carmen. Paradigms for abstracting systems. *Journal of Information Science*, 25(5), 1999, pp. 365-380.

Siddharthan, Advaith; Nenkova, Ani; McKeown, Kathleen. Syntactic Simplification for Improving Content Selection in Multi-Document Summarization. In: *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)*, pp. 896-902, Geneva, Switzerland, 2004.

Torres-Moreno, J.M; Velazquez-Morales, P.; Meunier, J.-G. Condensés de textes par des méthodes numériques. In : *Actes du colloque « L'analyse de données textuelles : De l'enquête aux corpus littéraires »*, Lexicometrica, 2004.