Sentence Reduction Algorithms to Improve Multi-document Summarization

Sara Botelho Silveira and António Branco

University of Lisbon, Portugal
Edifício C6, Departamento de Informática
Faculdade de Ciências, Universidade de Lisboa
Campo Grande, 1749-016 Lisboa, Portugal
{sara.silveira, antonio.branco}@di.fc.ul.pt,
WWW home page: http://nlx.di.fc.ul.pt/

Abstract. Multi-document summarization aims to create a single summary based on the information conveyed by a collection of texts. After the candidate sentences have been identified and ordered, it is time to select which will be included in the summary. In this paper, we describe an approach that uses sentence reduction, both lexical and syntactic, to help improve the compression step in the summarization process. Three different algorithms are proposed and discussed. Sentence reduction is performed by removing specific sentential constructions conveying information that can be considered to be less relevant to the general message of the summary. Thus, the rationale is that sentence reduction not only removes expendable information, but also makes room for further relevant data in a summary.

Keywords: Sentence reduction, Compression, Multi-document summarization

1 Introduction

The increased use of mobile devices brought concerns about text compression, by providing less space for the same amount of text. Compression must be accurate and all the information displayed should be essential. Multi-document text summarization seeks to identify the most relevant information in a collection of texts, complying with a compression rate that determines the length of the summary.

Ensuring at the same time the compression rate and the informativeness of the summary is not an easy task. The most common solution allows the last sentence to be cut in two in the number of words, where the exact compression rate has been reached, compromising the fluency and grammaticality of the summary, and thus the quality of the final text. An alternative is the one where the last candidate sentence is kept in full, surpassing the compression rate. None of these solutions is optimal. Compromising the compression rate by enhancing the quality of the text may not introduce relevant information. Still, compromising

the quality of the text can be troublesome for a user wanting to make use of the summary.

Given this, our proposal is to use sentence reduction to compress the extracted sentences down to their main content only, so that more information can fit into the summary, producing a more informative text. After the summarization process has determined the most significant sentences, sentential structures, that are less essential to figure in the summary's short space, can be removed.

The rationale behind using sentence reduction in a summarization context is twofold. On the one hand, it removes expendable information, generating a simpler and easier to read text. On the other hand, it allows the addition of more individual (reduced) sentences to the summary, that otherwise have not been included. Experiments made with human users [1] have shown that reduction indeed helps to improve the summaries produced.

Note that, sentence reduction is also referred in the literature as sentence compression. In this work, the expression "sentence reduction" is used to define "sentence compression", in order to distinguish it from "compression" itself. We name "compression" as the step that follows reduction in the summarization process, where the sentences identified as the most relevant ones are selected, based on a predefined compression rate, thus compressing the initial set of sentences contained in the collection of texts submitted as input.

At this point, consider the following list of sentences that can be part of the summary:

1. EU leaders signed a new treaty to control budgets on Friday.
2. Only Britain and the Czech Republic opted out of the pact, signed in Brussels at a summit of EU leaders.
3. UK Prime Minister David Cameron, who with the Czechs refused to sign it, said his proposals for cutting red tape and promoting business had been ignored.
4. The countries signed up to a promise to anchor in their constitutions – if possible – rules to stop their public deficits and debt spiralling out of control in the way that led to the eurozone crisis.
5. The treaty must now be ratified by the parliaments of the signatory countries.

This list contains 105 words. However, a compression rate of 80% of the original text states that the summary must only contain 84 words. As the sum of the words of the three first sentences (57 words) does not meet the desired total number of words for the summary, the fourth sentence is also added. Yet, by adding the fourth sentence, the summary makes up 92 words, so the total number of words defined by the compression rate has been surpassed in 9 words. The first option would be to cut the last nine words of the last sentence. That would produce an incorrect sentence.

There are particular constructions that can be removed from these sentences making room for the inclusion of more relevant information. Appositions, parenthetical phrases and relative clauses are examples of those constructions. Consider, for instance, the following expressions candidates for removal:

– The parenthetical phrase: *signed in Brussels at a summit of EU leaders*
– The relative clause: *who with the Czechs refused to sign*
– The parenthetical phrase: *if possible*

These expressions sum a total of 18 words. The last sentence that has not been added to the summary sums a total of 13 words. So, if all these expressions were removed from the sentences, we would have been able to include in the summary the last sentence. Otherwise that sentence would not be included in the final text, despite being relevant to the overall informativeness of the summary.

The summary, in which sentences have been simplified, contains 84 words and is shown below:

EU leaders signed a new treaty to control budgets on Friday. Only Britain and the Czech Republic opted out of the pact. UK Prime Minister David Cameron said his proposals for cutting red tape and promoting business had been ignored. The countries signed up to a promise to anchor in their constitutions rules to stop their public deficits and debt spiraling out of control in the way that led to the eurozone crisis. The treaty must now be ratified by the signatory countries' parliaments.

In a pilot study [2], Lin showed the potential of sentence reduction to improve a multi-document summarization system, using a noisy-channel model approach. Also, [3] used a machine learning approach to perform sentence extraction and compression for multi-document summarization, which proved to be effective in improving the quality of the summaries produced.

In a different perspective, [4] demonstrated that "a hybrid approach to sentence compression – explicitly modeling linguistic knowledge – rather than a fully data-driven approach" is the better way to perform sentence reduction.

As shown in the summary, it is possible to produce a summary containing the maximum relevant information conveyed by the original collection of texts. Hence, this summary can be a comprehensible and fluent one.

Thus, this work uses an hybrid approach by combining a statistical parser, that was trained on a specific corpus, with linguistic rules designed based on the output of the parser, defining the structure of the phenomena taken into account in this procedure.

Sentence reduction condenses, then, the initial summary, in order to produce a new text containing simpler, more precise and more concise sentences, and conveying only the essential information.

This paper is organized as follows: Section 2 reports the related work; Section 3 overviews the summarization process; Section 4 details the algorithms experimented in the context of sentence reduction; Section 5 describes a pilot study involving the three algorithms; Section 6 argues about the pros and cons of each algorithm; and, finally, in Section 7, some final conclusions are drawn.

2 Related Work

Text simplification is an Natural Language Processing (NLP) task that aims at making a text shorter and more readable by simplifying its sentences structurally, while preserving as much as possible the meaning of the original sentence. This task is commonly addressed in two ways: lexical and syntactic simplification. Lexical simplification involves replacing infrequent words by their simpler more common and accessible synonyms. Syntactic simplification, in turn, includes a linguistic analysis of the input texts, that produces detailed tree-structure representations, over which transformations can be made [5]. Syntactic simplification can also be named after sentence reduction.

Previous works ([6] and [7]) have focused on syntactic simplification, targeting specific types of structures identified using rules induced through an annotated aligned corpus of complex and simplified texts.

[8] used simplification in a single-document summarizer, by performing operations, based on the analysis of human abstracts, that remove inessential phrases from the sentences. [9] remove appositives and relative clauses in a preprocessing phase of a multi-document summarization process. Another proposal is the one of [10], that combine a simplification method, that uses shallow parsing to detect lexical cues that trigger phrase eliminations, with an HMM sentence selection approach, to create multi-document summaries.

Closer to our work is the work of [11], in which sentence simplification is applied together with summarization. However, they used simplification to improve content selection, that is, before extracting sentences to be summarized. Their simplification system is based on syntactic simplification performed using hand-crafted rules that specify relations between simplified sentences.

[12] applied sentence compression techniques to multi-document summarization, using a parse-and-trim approach to generate headlines for news stories. Constituents are removed iteratively from the sentence parse tree, using rules that perform lexical simplification – by replacing temporal expressions, preposed adjuncts, determiners, conjunctions, modal verbs –, and syntactic simplification – by selecting specific phenomena in the parse tree.

A different approach was used by [13], that experimented a tree-to-tree transduction method for sentence compression. They trained a model that uses a synchronous tree substitution grammar, which allows local distortions of a tree topology, used to capture structural mismatches between trees.

A word graph method, to create a single simplified sentence of a cluster of similar or related sentences, was used by [14]. Considering all the words in these related sentences, a directed word graph is built by linking word A to word B through an adjacency relation, in order to avoid redundancy. This method was used to avoid redundancy in the summaries produced.

[15] proposed a text summarization system that combines textual entailment techniques, to detect and remove information, with term frequency metrics used to identify the main topics in the collection of texts. In addition, a word graph method is used to compress and fuse information, in order to produce abstract summaries.

More recently, [16] investigated the usage of a machine translation technique to perform sentence simplification. They created a method for simplifying sentences by using Phrase Based Machine Translation, along with a re-ranking heuristic based on dissimilarity. Then, they trained it on a monolingual parallel corpus, and achieved state-of-the-art results.

Finally, [17] proposed new semantic constraints, to perform sentence compression. These constraints are based on semantic roles, in order to directly capture the relations between a predicate and its arguments.

3 Summarization process

The system used is an extractive multi-document summarizer that receives a collection of texts in Portuguese and produces highly informative summaries.

Summarization is performed by means of two main phases executed in sequence: clustering by similarity and clustering by keywords. Aiming to avoid redundancy, sentences are clustered by similarity, and only one sentence from each cluster is selected. Yet, the keyword clustering phase seeks to identify the most relevant content within the input texts. The keywords of the input texts are retrieved and the sentences that are successfully grouped to a keyword cluster are selected to be used in the next step of the process. Furthermore, each sentence has a score, which is computed using the *tf-idf* (term frequency – inverse document frequency) of each sentence word, smoothed by the number of words in the sentence. This score defines the relevance of each sentence and it is thus used to order all the sentences. Afterwards, the reduction process detailed in Section 4 is performed, producing the final summary. A detailed description of this extractive summarization process can be found in [18].

4 Sentence reduction

In this work, reduction is performed together with compression.

Firstly, from the original input list of sentences, a new list is created, by selecting one sentence at the time, until the total number of words in the list surpasses the maximum number of words determined by the compression rate.

Afterwards, sentences are reduced by removing the expendable information in view of the general summarization purpose. There are a number of structures that can be seen as containing "elaborative" information about the content already expressed.

Due to the fact that reduction removes words from the sentence, once sentences have been reduced, new sentences are added to the list of sentences to achieve the maximum number of words of the summary once again. Those newly added sentences are then reduced. This process is repeated while the list is changed or if the compression rate has not been meet.

Sentence reduction algorithms can consider many structures. These structures are described in Section 4.1. Afterwards, the algorithms that perform sentence reduction are discussed in Section 4.2.

4.1 Targeted structures

Different structures for different algorithms are targeted. At most six types of structures can be targeted:

- Appositions;
- Adjectives;
- Adverbs or adverb phrases;
- Parentheticals;
- Relative clauses;
- Prepositional phrases;

Appositions are noun phrases that describe, detail or modify its antecedent (also a noun phrase). The following sentence contains a an apposition (in bold).

ORIGINAL SENTENCE:
José Sócrates, ***primeiro-ministro****, e Jaime Gama querem cortar os salários dos seus gabinetes.*
José Sócrates, **the Prime Minister**, and Jaime Gama want to cut the salaries of their offices.
SIMPLIFIED SENTENCE:
José Sócrates e Jaime Gama querem cortar os salários dos seus gabinetes.
José Sócrates and Jaime Gama want to cut the salaries of their offices.

Adjectives qualify nouns or noun phrases, thus being structures prone to be removed. The following sentence contains an adjective (in bold).

ORIGINAL SENTENCE:
O palco tem um pilar ***central****, com 50 metros de altura.*
The stage has a **central** pillar, 50 meters high.
SIMPLIFIED SENTENCE:
O palco tem um pilar, com 50 metros de altura.
The stage has a pillar, 50 meters high.

Adverbs or adverb phrases are considered differently whether they appear in a noun or in a verb phrase, due to the usage of the adverbs of negation, which typically precede the verb. The adverbs appearing in a verb phrase are handled differently, to avoid removing negative adverbs and modifying the meaning of the sentence. The following sentence contains an adverb phrase (in bold).

ORIGINAL SENTENCE:
José Sócrates chegou ***um pouco*** *atrasado ao debate.*
José Sócrates arrived **a little late** to the debate.
SIMPLIFIED SENTENCE:
José Sócrates chegou atrasado ao debate.
José Sócrates arrived late to the debate.

Parenthetical phrases are phrases that explain or qualify other information being expressed. The following sentence contains a parenthetical phrase (in bold).

ORIGINAL SENTENCE:
O Parlamento aprovou, ***por ampla maioria****, a proposta.*
The Parliament approved **by large majority** the proposal.
SIMPLIFIED SENTENCE:
O Parlamento aprovou a proposta.
The Parliament approved the proposal.

Relative clauses are clauses that modify a noun phrase. They have the same structure as appositions, differing in the top node. The following sentence contains a relative clause (in bold).

ORIGINAL SENTENCE:
O Parlamento aprovou a proposta, ***que reduz os vencimentos dos deputados****.*
The Parliament approved the proposal, **which reduces the salaries of deputies**.
SIMPLIFIED SENTENCE:
O Parlamento aprovou a proposta.
The Parliament approved the proposal.

Prepositional phrases are phrases that modify nouns and verbs, indicating various relationships between subjects and verbs. They are used to include additional information within sentences. The following sentence contains a prepositional phrase (in bold).

ORIGINAL SENTENCE:
No Médio Oriente*, apenas Israel saudou a operação.*
In the Middle East, only Israel welcomed the operation.
SIMPLIFIED SENTENCE:
Apenas Israel saudou a operação.
Only Israel welcomed the operation.

In order to perform sentence reduction, a parse tree is created for each sentence, using a constituency parser for Portuguese [19]. The structures prone to be removed are identified in the tree using Tregex [20], a utility for matching patterns in trees. Tregex takes a parse tree and a regular expression pattern. It, then, returns a subtree of the initial tree which top node meets the pattern.

After identifying the subtrees representing each structure, these subtrees are replaced by null trees in the original sentence parse tree, removing its content and generating a new tree without the identified structure.

4.2 Algorithms

There were several algorithms that were experimented for sentence reduction. This section describes three of them: `main clause`, `blind removal`, and `best removal`. These algorithms differ not only in the structures that are removed, but also on the way those are removed. All these algorithms take a collection of sentences and return them reduced. The targeted structures are identified. Afterwards, reduced sentences are created by applying the algorithm that combines the removal of those structures.

The final step of the algorithm determines if the new reduced sentence can replace the former sentence, based on a specific criteria that takes into account the sentence score. In the summarization context, the sentence score defines the sentence relevance in the complete collection of sentences found in the input texts. This score is then a measure of informativeness. It states whether a sentence is important in the context of all the sentences in the texts to be summarized. Likewise, the score of a reduced sentence determines its informativeness.

The algorithms proposed in this work are described below. Thereafter, their pros and cons are discussed.

Main clause. This is a two step algorithm. First, the main clause of the sentence is identified. In this phase, other than the next, the desired subtree is selected, ignoring the other subtrees of the main tree. Consider the following sentence:

> *No Médio Oriente, apenas Israel saudou a operação.*
> In the Middle East, only Israel welcomed the operation.

The main clause of this sentence is:

> *Apenas Israel saudou a operação.*
> Only Israel welcomed the operation.

The expression "*No Médio Oriente*" is ignored, since it is not part of the main clause (in bold). The original sentence is replaced by the reduced one, in which further reduction rules are applied. In this step, clauses in a SVO structure are considered. If the sentence is not in this format, the whole original sentence is used.

After the main clause has been obtained, it is used to identify the structures to be removed. The subtrees of the structures are identified in the sentence parse tree. Five types of structures are targeted: appositions, adjectives, adverbs, parenthetical phrase, and relative clauses. In fact, this first step removes the previously mentioned prepositional phrases, since those are typically the structures used in a sentence before its main clause. So that, prepositional phrases are not taken into account in the next step. After the targeted passages have been identified, a reduced sentence is build by removing all the structures found in the main clause of the original sentence.

In this example, the main clause of this sentence contains just one removable passage:

– Adverb phrase – *Apenas* (only)

So that, the reduced sentence produced by this algorithm would be the following.

Israel saudou a operação.
Israel welcomed the operation.

A detailed description of this algorithm can be found in [21].

Blind removal. This algorithm takes four types of structures and removes them all from the original sentence. The structures considered in this algorithm are: appositions, parenthetical phrases, relative clauses, and prepositional phrases.

Consider the following sentence:

Também hoje, na conferência de líderes, o ministro dos Assuntos Parlamentares, Jorge Lacão, afirmou ter-se descoberto que o gabinete do primeiro-ministro tinha ficado de fora.
Today also, at the leadership conference, the Minister for Parliamentary Affairs, Jorge Lacão, said to have discovered that the office of the prime minister had been excluded.

Removable passages:

– Apposition – *Jorge Lacão*
– Prepositional phrase – *na conferência de líderes* (at the leadership conference)

In this algorithm, all these passages are removed from the original sentence, building the following reduced sentence.

Também hoje, o ministro dos Assuntos Parlamentares afirmou ter-se descoberto que o gabinete do primeiro-ministro tinha ficado de fora.
Today also, the Minister for Parliamentary Affairs said to have discovered that the office of the prime minister had been excluded.

This reduced sentence is the one used to be compared to the original sentence.

Best removal. This is an algorithm that uses the concept of power set, the set of all subsets of a given set. In the context of this work, the power set of a given sentence is composed by all the sentences obtained by combining the removal of the structures that have been identified as removable. Four types of structures are considered in this algorithm: appositions, parenthetical phrases, relative clauses, and prepositional phrases. Recall the sentence illustrated in the

previous algorithm and its removable passages.

> *Também hoje, na conferência de líderes, o ministro dos Assuntos Parlamentares, Jorge Lacão, afirmou ter-se descoberto que o gabinete do primeiro-ministro tinha ficado de fora.*
> Today also, at the leadership conference, the Minister for Parliamentary Affairs, Jorge Lacão, said to have discovered that the office of the prime minister had been excluded.

This sentence contains two removable passages (underlined): the apposition – *Jorge Lacão* –, and the prepositional phrase – *na conferência de líderes*.

The following example shows the original sentence and its score.

Também hoje, na conferência de líderes, o ministro dos Assuntos Parlamentares, Jorge Lacão, afirmou ter-se descoberto que o gabinete do primeiro-ministro tinha ficado de fora. Today also, at the leadership conference, the Minister for Parliamentary Affairs, Jorge Lacão, said to have discovered that the office of the prime minister had been excluded.	1.7200

The following table describes the sentences in the power set and their respective scores. These sentences were created by combining the removal of the identified structures. Their scores were obtained by summing the score (obtained in the summarization process) of each word composing the reduced sentence divided by the total number of words defining the new sentence. From the first sentence was removed the apposition phrase *Jorge Lacão*. The second sentence does not contain both removable passages *Jorge Lacão* and *na conferência de líderes*. Finally, the third one does not include the parenthetical phrase *na conferência de líderes*.

Também hoje, na conferência de líderes, o ministro dos Assuntos Parlamentares afirmou ter-se descoberto que o gabinete do primeiro-ministro tinha ficado de fora. Today also, at the leadership conference, the Minister for Parliamentary Affairs said to have discovered that the office of the prime minister had been excluded.	1.8175
Também hoje o ministro dos Assuntos Parlamentares afirmou ter-se descoberto que o gabinete do primeiro-ministro tinha ficado de fora. Today also the Minister for Parliamentary Affairs said to have discovered that the office of the prime minister had been excluded.	1.7053
Também hoje o ministro dos Assuntos Parlamentares, Jorge Lacão, afirmou ter-se descoberto que o gabinete do primeiro-ministro tinha ficado de fora. Today also the Minister for Parliamentary Affairs, Jorge Lacão, said to have discovered that the office of the prime minister had been excluded.	1.6000

After the power set has been defined, all the sentences are ordered by their score. As shown in the table, depending on the passage that has been removed or the combination of passages removed, the score of the reduced sentence keeps changing. This means that there are some expressions that contain more information than others, as the sentence score is a measure of informativeness. The reduced sentence will then be the sentence in the power set that has the maximum score.

4.3 Sentence selection

After the structures have been removed from the sentence, it is time to determine if this new reduced sentence should replace the original one.

Hence, the sentence score is considered. As mentioned above, in the summarization algorithm, the sentence score defines the sentence relevance to the complete collection of sentences obtained from the input texts. This score is computed using the *tf-idf* metric, which states that the relevance of a term not only depends on its frequency over the collection of texts, but also it depends on the number of documents in which the term occurs. Equation 1 describes the computation of the sentence score.

$$score_S = \frac{\sum_w tf - idf_w}{totalWords_S} \tag{1}$$

Hence, $score_S$ of the sentence S measures the relevance of this sentence considering the collection of sentences obtained from the input texts.

As words or expressions were removed from the original sentence to create the new reduced sentence, the score of this reduced sentence must be computed, considering only the words that it now contains. After having both sentence scores, the original sentence score is compared to the one of its reduced version. If the reduced sentence score is higher than the one of the original sentence, the reduced sentence replaces the former one in the summary.

This procedure ensures that sentence reduction indeed helps to improve the content of the summary, by including only the reduced sentences that contribute to maximize the informativeness of the final summary.

5 Pilot study

In order to illustrate the previous algorithms, a pilot study including a summary composed by two sentences has been conducted. Note that, in this study, after sentence reduction is applied to the summary, no more information is being added to it, despite that the summarization process completes the summary until the number of words defined by the compression rate is met.

Consider the following summary:

Esta foi a primeira pesquisa da série CNI/Ibope feita já com a lista oficial de candidatos à Presidência registrados no TSE (Tribunal Superior Eleitoral). Se a eleição fosse hoje, o presidente Luiz Inácio Lula da Silva, candidato à reeleição, teria 44% das intenções de voto, contra 25% de Geraldo Alckmin, de acordo com a pesquisa CNI/Ibope divulgada nesta sexta-feira.

This was the first survey in the series CNI/IBOPE, done already with the official list of presidential candidates registered in the TSE (Supreme Electoral Tribunal). If the election were today, President Luiz Inacio Lula da Silva, candidate for re-election, would have 44% of the vote, against 25% of Geraldo Alckmin, according to CNI/Ibope released on Friday.

This summary contains the following structures (underlined in the example) that can be targeted to be removed:

- Adverb#1 – *já*
- Adjective – *oficial*
- Parenthetical phrase – *Tribunal Superior Eleitoral*
- Adverb#2 – *hoje*
- Apposition phrase – *candidato à reeleição*
- Prepositional phrase – *de acordo com a pesquisa CNI/Ibope divulgada nesta sexta-feira.*

Also, the main clauses of each have been identified:

Main clause#1 *Esta foi a primeira pesquisa da série CNI/Ibope.*
Main clause#2 *O presidente Luiz Inácio Lula da Silva, candidato à reeleição, teria 44% das intenções de voto, contra 25% de Geraldo Alckmin, de acordo com a pesquisa CNI/Ibope divulgada nesta sexta-feira.*

Table 1 describes which structures were removed using each algorithm.

Table 1. Structures removed using each algorithm

	Main clause	Blind removal	Best removal
Adverb#1	N/A	-	-
Adjective	N/A	-	-
Parenthetical phrase	N/A	Yes	Yes
Adverb#2	N/A	-	-
Apposition phrase	Yes	Yes	Yes
Prepositional phrase	-	Yes	No
Main clause#1	Yes	-	-
Main clause#2	Yes	-	-

As illustrated in the previous Table, `main clause` does not take into account the first four structures, since its first step is to obtain the main clause, and those

structures are not part of the main clause. Yet, both `blind removal` and `best removal` do not consider adjectives and adverbs.

Table 2 describes the number of words removed by all these algorithms.

Table 2. Algorithm statistics (number of words removed)

	Main clause	Blind removal	Best removal
Sentence#1	16	3	3
Sentence#2	8	12	3
Total	24	15	6

In this very small example, there are some issues to be noticed. Despite that by applying `best removal` there is no more space in the summary for another sentence, it is possible to be sure that, with this algorithm, the best reduced sentence is created, maximizing the information of the current summary. Yet, `blind removal` removes all the structures allowing for more room to include new information, whether this information is relevant or not. Otherwise, when using `main clause`, two much information is lost, and there are no guarantees that the sentences added afterwards would include this information.

6 Discussion

The main assumption of a reduction process is that the identified structures are considered prone to be removed because they express additional information in the context of the sentence that can be avoided without jeopardizing the key content of the sentence they belong to. In addition, a well-defined sentence is easier to understand. Based on these two assumptions was created the very first approach to sentence reduction: the `main clause` algorithm. Firstly, the sentence is reduced to its main content, by identifying its SVO structure, and afterwards, the additional information is removed considering five types of passages.

However, this algorithm has some drawbacks. In fact, the SVO structure was difficult to retrieve, since there are many sentences that do not follow this structure. Furthermore, there were too many passages identified to be removed and sometimes the meaning of the sentence was not expressed, specially when adjectives and adverbs were removed.

These observations brought new decisions concerning the type of structures targeted. As not all these structures should be considered dispensable, a subset of them was selected. Considering their specific nature, appositions, parenthetical phrases, prepositional phrases and relative clauses are phrases that contain additional information to the content already expressed.

Thus, the next two algorithms, `blind removal` and `best removal`, considered only these types of passages. The next approach to the current reduction process, `blind removal`, defines that all the information expressed in those

structures is dispensable. Thus, all the candidate passages are blindly removed from all the sentences that go through this process. Considering the parenthetical nature of these passages, their simple removal would make room for more information to be included in the summary. In fact, the verification of the score, made after the sentence has been reduced, accounts for the informativeness of the sentence, and thus of the summary. However, after applying this algorithm, we concluded that there were some passages that by being removed would compromise the comprehensiveness of the text.

This conclusion drove the decision of applying the third algorithm, `best removal`. This algorithm aims to both maximize the information in the sentence and improve the comprehension of that sentence. By removing the structures carefully, taking into account the ones that improve the sentence informativeness, it is expected that consequently the informativeness of the summary also improves. Despite that by definition these structures constitute additional information, this information might not have been expressed yet in the summary. As stated above, the simple removal of all these structures can create incomprehensible sentences with too few information. The sentence score, by being the measure of the sentence informativeness, determines which of the reduced sentences created is the best, that is, the one that contains more information and, at the same time, discards the additional information.

In conclusion, `best removal` was then the final algorithm selected, since it verifies three important conditions: (1) it considers only the structures that indeed make up additional information; (2) it produces the best combination of a reduced sentence, and (3) in itself it takes into account the informativeness of the reduced sentences within the whole collection of sentences by considering their score.

7 Final remarks

This paper presents three possible algorithms to perform sentence reduction. The idea behind all these algorithms is detailed. Also, pros and cons of each one are commented and some final conclusions about their differences are drawn.

The approach that combines summarization with sentence reduction is an effective procedure that seeks to maximize the relevant information within a summary. In fact, by reducing the sentences from the initial set of sentences into their main content, sentence reduction allows for the inclusion of further sentences containing novel and relevant information. Moreover, the type of structures that are removed is also a matter of concern. As was discussed above, there are some structures that should not be removed, in order to ensure that the meaning of a sentence is kept. The algorithms presented also take this issue into account.

In the context of summarization, such a combination – summarization followed by sentence reduction – aims to produce highly informative summaries, containing the maximum amount of significant information.

References

1. Silveira, S.B., Branco, A.: Enhancing multi-document summaries with sentence simplification. In: ICAI 2012: International Conference on Artificial Intelligence, Las Vegas, USA (July 2012) 742–748
2. Lin, C.Y.: Improving summarization performance by sentence compression: a pilot study. In: Proceedings of the sixth international workshop on Information retrieval with Asian languages - Volume 11. AsianIR '03, Stroudsburg, PA, USA, Association for Computational Linguistics (2003) 1–8
3. Berg-Kirkpatrick, T., Gillick, D., Klein, D.: Jointly learning to extract and compress. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1. HLT '11, Stroudsburg, PA, USA, Association for Computational Linguistics (2011) 481–490
4. Marsi, E., Krahmer, E., Hendrickx, I., Daelemans, W.: Empirical methods in natural language generation. Springer-Verlag, Berlin, Heidelberg (2010) 45–66
5. Feng, L.: Text simplification: A survey. Technical report, The City University of New York (2008)
6. Chandrasekar, R., Doran, C., Srinivas, B.: Motivations and methods for text simplification. In: In Proceedings of the Sixteenth International Conference on Computational Linguistics (COLING '96). (1996) 1041–1044
7. Jing, H.: Sentence reduction for automatic text summarization. In: Proceedings of the sixth conference on Applied natural language processing, Morristown, NJ, USA, Association for Computational Linguistics (2000) 310–315
8. Jing, H., McKeown, K.R.: Cut and paste based text summarization. In: Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference. NAACL 2000, Stroudsburg, PA, USA, ACL (2000) 178–185
9. Blair-Goldensohn, S., Evans, D., Hatzivassiloglou, V., Mckeown, K., Nenkova, A., Passonneau, R., Schiffman, B., Schlaikjer, A., Advaith, Siddharthan, A., Siegelman, S.: Columbia university at duc 2004. In: Proceedings of the 2004 document understanding conference (DUC 2004). HLT/NAACL 2004, Boston, Massachusetts (2004) 23–30
10. Conroy, J., Schlesinger, J., Stewart, J.: Classy query-based multidocument summarization. In: Proceedings of 2005 Document Understanding Conference, Vancouver, BC (2005)
11. Siddharthan, A., Nenkova, A., McKeown, K.: Syntactic simplification for improving content selection in multi-document summarization. In: COLING '04: Proceedings of the 20th international conference on Computational Linguistics, Morristown, NJ, USA, ACL (2004) 896
12. Zajic, D., Dorr, B.J., Lin, J., Schwartz, R.: Multi-candidate reduction: Sentence compression as a tool for document summarization tasks. Inf. Process. Manage. **43**(6) (2007) 1549–1570
13. Cohn, T., Lapata, M.: Sentence compression as tree transduction. J. Artif. Intell. Res. (JAIR) **34** (2009) 637–674
14. Filippova, K.: Multi-sentence compression: finding shortest paths in word graphs. In: Proceedings of the 23rd International Conference on Computational Linguistics. COLING '10, Stroudsburg, PA, USA, ACL (2010) 322–330
15. Lloret, E.: Text Summarisation based on Human Language Technologies and its Applications. PhD thesis, Universidad de Alicante (2011)
16. Wubben, S., van den Bosch, A., Krahmer, E.: Sentence simplification by monolingual machine translation. In: ACL – The 50th Annual Meeting of the Association

for Computational Linguistics, Proceedings of the Conference, July 8-14, 2012, Jeju Island, Korea - Volume 1: Long Papers, The Association for Computer Linguistics (2012) 1015–1024
17. Yoshikawa, K., Iida, R., Hirao, T., Okumura, M.: Sentence compression with semantic role constraints. In: ACL – The 50th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, July 8-14, 2012, Jeju Island, Korea - Volume 2: Short Papers, The Association for Computer Linguistics (2012) 349–353
18. Silveira, S.B., Branco, A.: Combining a double clustering approach with sentence simplification to produce highly informative multi-document summaries. In: IRI 2012: 14th International Conference on Artificial Intelligence, Las Vegas, USA (August 2012) 482–489
19. Silva, J., Branco, A., Castro, S., Reis, R.: Out-of-the-box robust parsing of Portuguese. In: Proceedings of the 9th Encontro para o Processamento Computacional da Língua Portuguesa Escrita e Falada (PROPOR). (2010) 75–85
20. Levy, R., Andrew, G.: Tregex and Tsurgeon: Tools for querying and manipulating tree data structures. In: Proceedings of the 5th Language Resources and Evaluation Conference (LREC). (2006)
21. Silveira, S.B., Branco, A.: Compressing multi-document summaries through sentence simplification. In: ICAART 2013: 5th International Conference on Agents and Artificial Intelligence, Barcelona, Spain (February 2013)