Enhancing Multi-document Summaries with Sentence Simplification

Sara Botelho Silveira and António Branco
University of Lisbon
Departamento de Informática
Faculdade de Ciências, Universidade de Lisboa
1749-016 Lisboa, Portugal
{sara.silveira, antonio.branco}@di.fc.ul.pt

Abstract— *This paper investigates the usage of a sentence simplification module that aims at improving automatically generated summaries built from a collection of texts. We describe an experiment with human subjects, that was conducted to assess the actual impact of sentence simplification in enhancing the summary quality. Motivated by those results, we present an automatic sentence simplification module that removes, from a sentence, specific sentential constructions that do not add critical information to the general message of that sentence. The rationale is that sentence simplification not only removes expendable information, but also makes room in a summary for further relevant data.*

Keywords: Multi-document summarization, sentence simplification, text simplification

1. Introduction

Text simplification is a NLP task that aims at making a text shorter and more readable, by simplifying its sentences structurally, into shorter and simpler sentences, while preserving the meaning and the information of the original sentence as much as possible. This task can be addressed in three ways: lexical, syntactic and discourse simplification. Lexical simplification involves replacing infrequent words by their simpler synonyms. To perform syntactic simplification, a linguistic representation of the text is produced as a tree-structure over which transformations are made [1]. Discourse simplification is concerned with maintaining coherence and cohesion of the simplified text: while syntactic simplification is applied to one sentence at a time, discourse simplification considers the interactions across sentences.

In order to improve the quality of the summary delivered to the end user, we discuss the inclusion of a syntactic simplification module within an extractive summarization system for Portuguese. Sentences are simplified after summarization to produce highly informative summaries.

Previous works ([2] and [3]) have focused mainly on syntactic simplification. Several types of structures are removed from a sentence based on rules induced using an annotated aligned corpus of complex and simplified texts. These structures include, for instance, passages delimited by punctuation, subordination and coordinating conjunctions, relative pronouns and boundaries of clauses and phrases. Closer to our work is the work of [4], in which simplification is used to improve content selection in a summarization system, that is, before summarizing. It is a syntactic simplification system, that uses hand-crafted rules which specify relations between simplified sentences.

A text usually contains passages that seek to explain or qualify other phrases, by providing background information about entities, or relating those entities to the discourse. These passages are parenthetical phrases. Appositions are a specific type of parentheticals, composed by a noun phrase that describes, details or modifies its antecedent (also a noun phrase). This type of phrases often add no crucial content to the correct comprehension of the sentence. Thus, we intend to remove those passages affecting the content expressed in the text to be summarized as less as possible.

The following examples present two sentences, from which parenthetical and apposition phrases (in bold) were removed, creating the corresponding simplified sentences.

Parenthetical phrase removal:

ORIGINAL SENTENCE:

Vilaça não duvidou declarar que Jones Bule ***(como ele chamava ao inglês)*** *fizera do Ramalhete "um museu".*

"Vilaça was the first to declare that "Jones Bule" **(as he called the Englishman)** had made of Ramalhete "a veritable museum"."

SIMPLIFIED SENTENCE:

Vilaça não duvidou declarar que Jones Bule fizera do Ramalhete "um museu".

"Vilaça was the first to declare that "Jones Bule" had made of Ramalhete "a veritable museum"."

Apposition phrase removal:

ORIGINAL SENTENCE:

Em 1858, Monsenhor Buccarini, ***Núncio de Sua Santidade,*** *visitara-o com ideia de instalar lá a Nunciatura.*

"In 1858, Monsignor Buccarini**, the Papal Nuncio,** had visited it with a view to establishing his residence there."

SIMPLIFIED SENTENCE:

Em 1858, Monsenhor Buccarini visitara-o com ideia de instalar lá a Nunciatura.

"In 1858, Monsignor Buccarini had visited it with a view to establishing his residence there."

To assess the possible improvements that simplification may bring to a summary, we performed a test with language experts, by asking them to rate simplified summaries over summaries whose sentences had not been simplified. The experts stated that simplification actually improves the final summary, since it allows the creation of a more simpler and incisive text, preserving at the same time the information to be conveyed. Based on these conclusions, we developed an automatic sentence simplification system that is included in a multi-document summarization system.

This paper is organized as follows: Section 2 describes the manual experiments; Section 3 describes the simplification system; Section 4 presents an evaluation of the summarization system with and without the simplification module; and, finally, in Section 5, some conclusions are drawn.

2. Manual Experiment

Experiments with human experts were made to assess whether text simplification improves the quality of a summary. Source texts were randomly selected from *TeMário* corpus [5]. Automatic summaries were build using GistSumm [6], a single-document summarizer for Portuguese (the only one available on-line). These summaries have a compression rate of 50%, since texts with a higher compression rate will not have enough material to work with. Then, simplified summaries were constructed by hand, from the automatic summaries, by removing parenthetical and apposition phrases from their sentences. This procedure was manually performed and aimed at simulating the process an automatic text simplification tool would execute.

In the first experiment, three language experts[1] were asked to perform two different tasks, in which they considered different texts (the source texts, the automatic and the simplified summaries). Also, they were invited to comment not only the tasks, but also the overall quality of the summaries. This experiment was based on two goals, Goal#1 and Goal#2, which will be pursued in two different tasks, Task#1 and Task#2, respectively. The two tasks, their goals and their evaluation are described in the following sections.

2.1 Task#1

This task aimed at verifying whether the simplified summary was a good summary for the original text (without taking into account the automatic summary). In order to perform this task, five texts, one from each section of the corpus (Special, World, Opinion, International, Politics) were selected. Experts were given the same five texts and their respective simplified summaries. Afterwards, they rated (0, bad – 5, very good) the simplified summaries concerning the original texts.

[1]Three researchers on Language Technology, graduated in linguistics.

Goal#1:

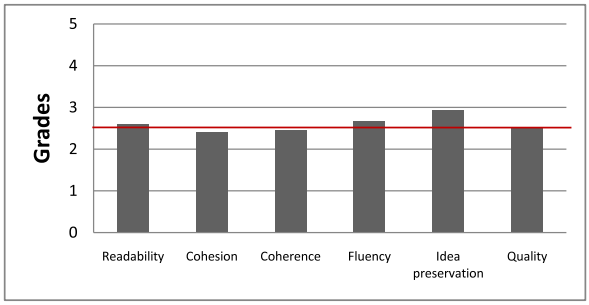
Confront the original text with a simplified summary, in order to assess the summary's quality.

Task#1:

Rate the **simplified summary** over the **original text**, concerning:
1. Readability (0 – 5);
2. Cohesion (0 – 5);
3. Coherence (0 – 5);
4. Fluency (0 – 5);
5. Original text idea preservation (0 – 5);
6. Overall quality (0 – 5);

Results are illustrated in Figure 1. By observing the graph, we can conclude that by having a low but positive grade (2.53) the *overall quality* of the summaries is acceptable. It is important to note that the best parameter is the *original text idea preservation* (2.93). Thus, we can infer that even though the simplified summary has less content, it does not mean that it does not maintain enough information conveyed by the original text. However, the way the information is presented is not the best, meaning the *fluency* (2.67) and the *readability* (2.6) of the text can be improved. The worst evaluated parameters are *coherence* (2.47) and *cohesion* (2.4). According to the experts, the sentence selection causes a lack of cohesion in the overall text. Since sentences seem disconnected and not linked with each other, text coherence is also compromised.

Fig. 1: Simplified summaries features.



2.2 Task#2

This task aimed to confront both summaries to verify if the simplified summary, though with less content, can still convey the same information within the automatic one. It comprises two phases. In PHASE I, the simplified summary is rated vis a vis the automatic summary, considering the same set of features used in Task#1. PHASE II seeks to understand if the simplification process would not remove information that is relevant to the overall understanding of

the text. Thus, experts have answered a survey that compares the automatic and the simplified summaries.

Goal#2:

Compare an automatic summary with a simplified summary in order to assess the simplified summaries quality.

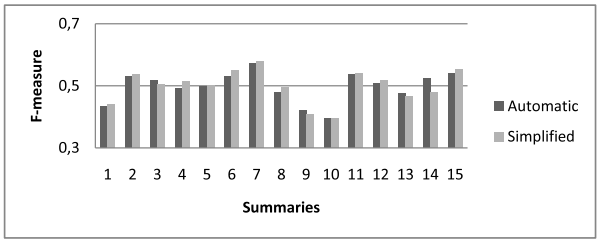
Task#2:

PHASE I: Rate the **simplified summary**.
1. Readability (0 – 5);
2. Cohesion (0 – 5);
3. Coherence (0 – 5);
4. Fluency (0 – 5);
5. Original text idea preservation (0 – 5);
6. Overall quality (0 – 5);

PHASE II: Answer the following questions by comparing the **simplified summary** with the **automatic summary**:
1. Which of the summaries is the best?
2. Consider that the simplified summary was built over the automatic one, by removing expressions considered unnecessary:
2.1. Were the removed expressions necessary?
2.2. Are there any expressions that you consider that should not have been removed?
2.3. Does the sentences context remain the same?
2.4. Does the simplified summary preserve the same idea of the original one?
2.5. Can the simplified summary replace the automatic one without losing information?
3. Other comments.

Before giving the two types of summaries to be evaluated by the human experts, both summaries were compared using ROUGE [7]. The graph in Figure 2 shows the f-measure values for all the 15 summaries used in this experiment. There is a slightly difference between them, as the average of the f-measure values for both automatic summaries (0.49752) and simplified summaries (0.49826) confirms.

Fig. 2: ROUGE f-measure values for automatic and simplified summaries.

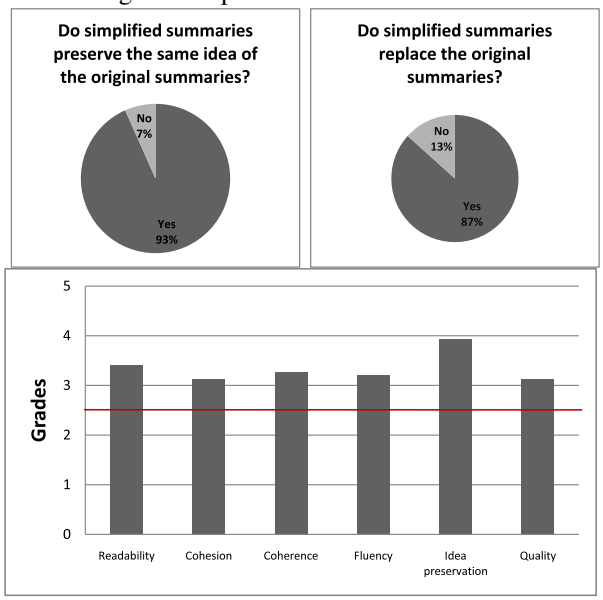


By comparing the graph in Figure 2 with the results in Figure 3, we can conclude that ROUGE does not reflect all the modifications performed in the simplified summaries, evidencing the need for manual evaluation. Nevertheless, it is a very useful tool to perform evaluation during system development, as metrics are obtained automatically.

To actually execute Task#2, each of the three experts was given five different texts. One text from each section of the corpus. Thus, we evaluated 15 texts, 3 from each section. This way, different sorts of texts were tested. Despite all of them being news texts, different sections may have different vocabulary and different writing styles. In PHASE I, experts did not have access to the original texts, so their conclusions were only based in the information contained in both summaries. Still, the automatic summary, by being longer and by having more content, should be expected to have more information than the simplified summary.

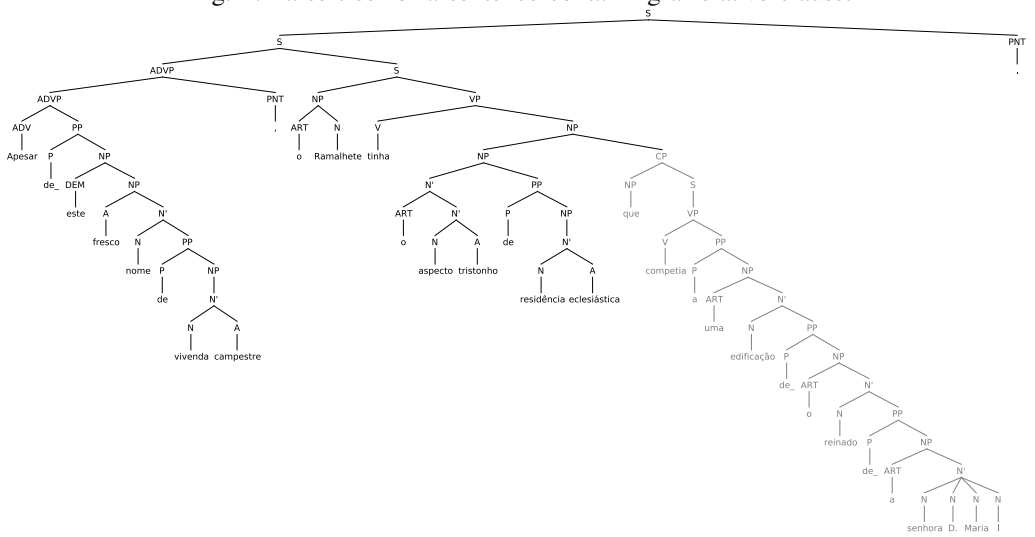
Figure 3 presents the answers given by the experts to the survey.

Fig. 3: Simplified summaries features.



The answers to question 1. were not conclusive, since, as experts observed, the differences between automatic and simplified summaries are minimal. However, they definitely stated that simplified summaries can replace the non-simplified ones, which means that no relevant information is being missed. Thus, the information that was removed is not essential for the comprehension of the text. In addition, experts reported that the simplified summaries were simpler than the automatic ones. Also, they stated that the majority of the removed expressions were not necessary to the comprehension of the text content. Finally, they mentioned that, mainly when long parenthetical information is removed, text readability is improved.

Fig. 4: Parse tree for a sentence containing a relative clause.



Since very similar texts were being compared, an improvement in the overall quantitative grades was expected. Nevertheless, the *original text idea preservation* was still the best feature (3.93), which lead to the fact that the information removed was indeed not essential in the summary. Moreover, this removal did not have a significant impact on text *readability* (3.4), seeing that the simplified text remains readable despite the modifications. This is not only reflected in the text *coherence* (3.27), *cohesion* (3.2), and *fluency* (3.13), but mainly in the text *overall quality* (3.13).

With this experiment, we can conclude that the simplified summaries preserve the original text idea. Also, the relevant information conveyed by the original text is maintained, allowing at the same time an optimization of the summary compression rate.

3. Simplifier

Based on the promising conclusions driven from the manual experiments, this task has been performed automatically. The simplifier is a module integrated in a multi-document summarization system, which produces generic summaries from a collection of texts in Portuguese, from any domain. After selecting the relevant content to be part of the summary, the chosen sentences are simplified to produce highly informative summaries. A detailed description of this summarizer can be found in [8].

The simplifier receives a set of sentences and retrieves them simplified. It includes two main phases, analysis and transformation, which are described in the next sections.

3.1 Analysis

The analysis phase aims to identify in each sentence the expressions which are candidate for removal. Three types of structures are recognized: (1) relative clauses; (2) parentheticals – explanatory or qualifying phrases; and (3) appositions – specific type of parentheticals, composed by a noun phrase that describes, details or modifies its antecedent (also a noun phrase). These phrases are candidates to removal, since they may introduce extra information whose removal does not affect the information eventually conveyed too much.

In order to identify the removable passages, a constituency parser for Portuguese [9] is used to build the sentence parse tree. Then, for each sentence, relative clauses, parentheticals and appositions are retrieved.

The following sentence contains a **relative clause**:

Apesar deste fresco nome de vivenda campestre, o Ramalhete tinha o aspecto tristonho de residência eclesiástica que competia a uma edificação do reinado da senhora D. Maria I.

"Despite that fresh green name worthy of some rural retreat, Ramalhete, as befitted a building dating from the reign of Queen Maria I, had the gloomy appearance of an ecclesiastical residence."

S *Apesar deste fresco nome de vivenda campestre, o Ramalhete tinha o aspecto tristonho de residência eclesiástica.*

"Despite that fresh green name worthy of some rural retreat, Ramalhete had the gloomy appearance of an ecclesiastical residence."

CP *que competia a uma edificação do reinado da senhora D. Maria I*

"as befitted a building dating from the reign of Queen Maria I"

Figure 4 illustrates the sentence parse tree in which both clauses are identified.

The first clause is the main one (starting in the top node *S* in the parse tree), and the second one is the subordinate relative clause (starting in the top node *CP* in the parse tree). Subtrees starting with the tag `CP` (complementizer phrase) are stored to be processed later.

Fig. 5: Parse tree for a sentence containing an apposition phrase.

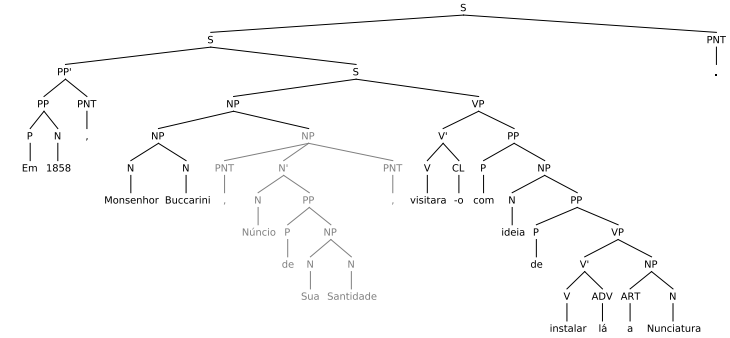
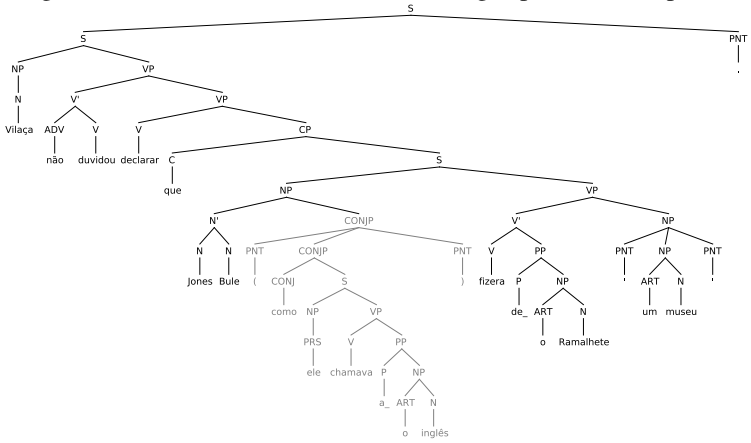


Fig. 6: Parse tree for a sentence containing a parenthetical phrase.



Appositions (specific type of parentheticals) are noun phrases, enclosed by commas or dashes, that define, describe or modify their antecedent which is also a noun phrase. Figure 5 depicts the parse tree for the following sentence, which includes an apposition:

Em 1858, Monsenhor Buccarini, ***Núncio de Sua Santidade,*** *visitara-o com ideia de instalar lá a Nunciatura.*

"In 1858, Monsignor Buccarini**, the Papal Nuncio,** had visited it with a view to establishing his residence there."

Considering the sentence parse tree, subtrees that start with the tags `NP` (noun phrase) or `AP` (adjective phrase), and whose leftmost child is a dash or a comma and whose rightmost child is a matching dash or comma or a period, are defined as appositions.

Parentheticals are phrases, enclosed either by parenthesis, or by commas or dashes, that explain or qualify other information being expressed. Consider the following sentence containing a parenthetical phrase:

Vilaça não duvidou declarar que Jones Bule ***(como ele chamava ao inglês)*** *fizera do Ramalhete "um museu".*

"Vilaça was the first to declare that "Jones Bule" **(as he called the Englishman)** had made of Ramalhete "a veritable museum"."

Figure 6 illustrates the parse tree for this sentence. In the parse tree, parentheticals are defined by a node, represented by any tag, which leftmost child is the opening punctuation token and its rightmost child is a closing punctuation token of the same type.

3.2 Transformation

The transformation phase aims to decide which of the phrases, retrieved in the analysis phase, can be removed without impacting on the sentence's overall informativity. Consider the following sentence:

Na conferência de líderes de ontem, a primeira da nova sessão, que a vice-presidente da bancada do PS, Ana Catarina Mendes, disse esperar que fique marcada pela 'governabilidade', ficou estabelecido que a execução orçamental será o tema central da Comissão Permanente.

"At yesterday's Leaders' Conference, the first of the new session which PS' delegation vice-president, Ana Catarina Mendes, said to hope it would be remembered by 'governability', it was settled that budget execution will be the main subject of the Standing Committee."

This sentence for instance has three removal candidates: a relative clause – *"que a vice-presidente da bancada do PS, Ana Catarina Mendes, disse esperar que fique marcada pela 'governabilidade"'*; and two appositions – *"a primeira da nova sessão"* and *"Ana Catarina Mendes"*.

For each sentence, the algorithm first considers the relative clauses, then the apposition phrases and finally the parenthetical ones. In order to obtain the new sentence, the parse tree of the original sentence is traversed until the index of the first leaf of the phrase to be removed is found. Also, the index of the last leaf of the phrase is gathered. Then, the parent node of these leaves is retrieved and the subtree is removed. The new sentence is then created without the subtree representing the phrase removed.

Afterwards, the simplification score is computed for both the main sentence and the new sentence. It is important to note that the simplifier is part of a text summarization system. This system computes the scores of the words composing the collection of sentences submitted as input, in order to select the most relevant sentences within that collection. Thus, the simplifier uses the simplification score to compare simplified sentences. This score is computed as the sum of each word score[2] composing the sentence divided by the total number of words of the sentence.

$$simplificationScore_{sentence} = \frac{\sum_{word \in sentence} score_{word}}{totalWords_{sentence}}$$

Considering the previous sentence and its removal candidates, the new sentences obtained along their simplification scores are shown in Table 1.

Table 1: Candidates to replacement.

	Candidate sentences	Score
1	*Na conferência de líderes de ontem, a primeira da nova sessão, que a vice-presidente da bancada do PS, Ana Catarina Mendes, disse esperar que fique marcada pela 'governabilidade', ficou estabelecido que a execução orçamental será o tema central da Comissão Permanente.* "At yesterday's Leaders' Conference, the first of the new session which PS' delegation vice-president, Ana Catarina Mendes, said to hope it would be remembered by 'governability', it was settled that budget execution will be the main subject of the Standing Committee."	0.01045
2	*Na conferência de líderes de ontem, a primeira da nova sessão, ficou estabelecido que a execução orçamental será o tema central da Comissão Permanente.* "At yesterday's Leaders' Conference, the first of the new session, it was settled that budget execution will be the main subject of the Standing Committee."	0.01615
3	*Na conferência de líderes de ontem, ficou estabelecido que a execução orçamental será o tema central da Comissão Permanente.* "At yesterday's Leaders' Conference, it was settled that budget execution will be the main subject of the Standing Committee."	0.01926

The first sentence in the Table is the original sentence and its simplification score is also compared with the ones of the simplified sentences. As the algorithm first considers the relative clauses, none of the simplified sentences contain the relative clause (*que a vice-presidente da bancada do PS, Ana Catarina Mendes, disse esperar que fique marcada pela 'governabilidade'*). Thus, the second simplified sentence was built by eliminating the relative clause. The third one is obtained by removing not only the relative clause, but also the first apposition referred (*a primeira da nova sessão*). The other apposition (*Ana Catarina Mendes*) is never considered since it is included in the relative clause. Owing that the third simplified sentence has the highest simplification score, it is the one selected to replace the original sentence in the final summary.

The same algorithm is executed for all the sentences which are candidates to be in the summary, ensuring the creation of a concise and highly informative text.

4. Evaluation

In order to perform evaluation, the *CSTNews* corpora [10], an annotated corpus composed of texts in Portuguese, was used. It contains 50 sets of news texts from several domains, for a total of 140 documents, 2,247 sentences, and 47,350 words. Each set contains, on average, 3 documents, which address the same subject, and an ideal summary (built manually). Ideal summaries have an average of 137 words, resulting in an average compression rate of 85%.

First, for each set of *CSTNews*, two types of summaries were created using our summarizer: simplified and non-simplified summaries. Simplified summaries are built by executing the whole summarization process. Otherwise, non-simplified ones are created without running the simplification module. In addition, summaries were generated using `GistSumm`[3] to serve as a baseline for our work. Summaries were built using a compression rate of 85% (average compression rate of the ideal summaries), meaning that the summary contains 15% of the words of the set of texts.

Afterwards, ROUGE [7] was used to derive precision, recall and f-measure metrics for the automatic summaries. ROUGE-L (longest common subsequence) was the metric selected, since it identifies the common subsequences between two sequences. This is a fairer metric compared to the original ROUGE-N metric, due to the fact that the simplification process introduces gaps in the extracted sentences. These gaps would not be taken into account when using a metric that considers co-occurring matches. Table 2 presents the results obtained.

Table 2: ROUGE-L evaluation metrics.

	GistSumm	Non-simplified	Simplified
Precision	0.3847	0.4276	0.4364
Recall	0.4362	0.4854	0.4942
F-Measure	0.4040	0.4492	0.4585

[2]The word score is the `tf-idf` score of the word considering the collection of texts to be summarized.

[3]Despite being a single-document summarizer, `GistSumm` also builds multi-document summaries by means of an option in its interface.

The complete summarization process has an overall better performance than the baseline considered, since both results with simplified and non-simplified summaries overcome the results obtained by `GistSumm`. The recall values obtained by our summarizer are very encouraging. These values indicate that there is a higher density of words that are both in the automatic summaries and in the ideal summaries. Retrieving the most relevant information in a sentence by discarding the less relevant data ensures that the summary indeed contains the most important information conveyed. This is a direct result of the simplification process. The precision values, by being identical, suggest that all summaries cover nearly the same topics. Intuitively, the precision values for the simplified summaries should decrease, since less in-sequence matches are likely to be found between them and the ideal summaries. When computing the f-measure value, by combining both precision and recall, we can confirm that both simplified and non-simplified summaries are better than `GistSumm` summaries.

Finally, note that the simplified summaries achieve better results than the non-simplified summaries. Despite being small, this difference between both summaries is very interesting. Yet, the simplification process is reflected in the evaluation of the summaries, implying that a simplification process, containing more sophisticated rules, can indeed be expected to help produce even better summaries.

5. Conclusions

This paper presents a text simplification system which is part of a text summarization system. The simplification system is focused on improving the summarization output, by simplifying the sentences composing the summaries. Our premise is that simplified sentences contain as much information as needed, helping to create more informative texts.

As manual experiments have hinted out, the simplification system is a huge contribution to the summarizer. Three main reasons lead to the decision of applying simplification after summarization: (1) this way, no relevant information is excluded before selecting the sentences to include in the summary, once all the computation regarding each original sentence has been done; (2) the simplification algorithm described does not remove the identified structures without constraints, since we aim to ensure that no crucial content is deleted from the sentence; (3) during the compression process, it is possible to add more relevant content to the summary that was not being considered in the initial list of sentence candidates.

Considering that most of the experiments have been done for the English language, we intended to prove that simplification actually improves summarization of texts in Portuguese, by using a corpora of Portuguese texts in all the experiments made.

References

[1] L. Feng, "Text simplification: A survey," The City University of New York, Tech. Rep., 2008.

[2] R. Chandrasekar, C. Doran, and B. Srinivas, "Motivations and methods for text simplification," in *In Proceedings of the Sixteenth International Conference on Computational Linguistics (COLING '96)*, 1996, pp. 1041–1044.

[3] H. Jing, "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, pp. 310–315.

[4] A. Siddharthan, A. Nenkova, and K. McKeown, "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, p. 896.

[5] L. Rino and T. Pardo, "Temário: Um corpus para a sumarização automática de textos," São Carlos – SP, Tech. Rep., 2003.

[6] T. A. S. Pardo, L. H. M. Rino, and M. das Graças Volpe Nunes, "Gistsumm: A summarization tool based on a new extractive method." in *PROPOR*, ser. Lecture Notes in Computer Science. Springer, 2003, pp. 210–218.

[7] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, S. S. Marie-Francine Moens, Ed. Barcelona, Spain: ACL, July 2004, pp. 74–81.

[8] S. B. Silveira and A. Branco, "Extracting multi-document summaries with a double clustering approach," in *Proceedings of the 17th International Conference on Applications of Natural Language Processing to Information Systems*. Groningen, The Netherlands: Springer, June 2012.

[9] J. Silva, A. Branco, S. Castro, and R. Reis, "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, pp. 75–85.

[10] P. Aleixo and T. A. S. Pardo, "Cstnews: Um córpus de textos jornalísticos anotados segundo a teoria discursiva multidocumento cst (cross-document structure theory)." Universidade de São Paulo, Tech. Rep., 2008.