# SIMBA: An Extractive Multi-document Summarization System for Portuguese

Sara Botelho Silveira and António Branco

Department of Informatics
Faculty of Sciences, University of Lisbon
sara.silveira@di.fc.ul.pt, antonio.branco@di.fc.ul.pt
WWW home page: http://nlx.di.fc.ul.pt

**Abstract.** This is a proposal for demonstration of SIMBA in PROPOR 2012. SIMBA is an extractive multi-document summarization system that aims at producing generic summaries guided by a compression rate defined by the user. It uses a double-clustering approach to find the relevant information in a set of texts. In addition, SIMBA uses a sentence simplification procedure as a mean to ensure summary compression. Furthermore, simplification seeks to produce simpler and more incisive sentences, containing the amount of information required.

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

## 1 Introduction

This document is a proposal for the demonstration of SIMBA, an extractive multi-document summarizer for the Portuguese language.

The present document is organized as follows: Section 2 provides a summarized description of SIMBA and Section 3 describes the demo presentation.

## 2 System overview

SIMBA receives a collection of Portuguese texts, from any domain, and produces informative summaries, for a generic audience. The summary length is determined by the compression rate value submitted by the user. A detailed description of SIMBA can be found in [3]. Summarization is performed by executing five main stages – identification, matching, filtering, reduction, and presentation – briefly described in the following subsections.

### 2.1 Identification

The identification stage annotates the texts, submitted by the user. A set of shallow processing tools for Portuguese, LX-Suite [1], identifies sentence and paragraph boundaries and tag words with its corresponding POS and lemmata. Also, a parse tree representing the sentence syntactic structure is built, using LX-Parser [2]. Henceforth, the collection of texts is handled as a set of sentences.

### 2.2 Matching

The matching stage aims at identifying relevant information in the collection of texts. This stage executes three steps: (1) computing sentence score; (2) clustering sentences by similarity; and (3) clustering sentences by keywords.

The sentence main score is computed as the sum of the `tf-idf` score (computed considering the lemma) of each of its words. Beyond the sentence main score, each sentence has an extra score value, which is used in the summarization process to reward or penalize the sentences, by adding or removing predefined score values. The sentence score is thus the sum of these two scores.

The next step of the mapping process seeks to identify redundant sentences, by clustering the sentences considering their degree of similarity. A similarity value is checked against with a predefined threshold to determine whether those sentences convey the same information or not. All sentences in the collection of texts are considered. Sentence-to-sentence similarity is computed. If the similarity degree is above the threshold, the sentences are grouped in the same cluster. From each cluster, the sentence with the highest score is selected to the next phase, and is rewarded accordingly through the addition of an extra score value.

SIMBA produces a generic summary, so it is not focused on a specific matter. Thus, the keywords that represent the global topic within the collection of texts are identified. Each keyword represents a cluster. Each sentence is added to the cluster represented by the keyword that occurs more often in that sentence. The sentences that have indeed be clustered are the ones used in the next stage, and are rewarded through the addition of an extra score value. The ones that have not been clustered are not considered to be relevant in the global context, so that these sentences are discarded in this phase.

### 2.3 Filtering

In this phase, sentences with less than ten words are penalized, and the ones with more than ten words are assigned with an extra score value. Afterwards, sentences are ranked by their complete score, defining the order of the sentences that can be chosen to be part of the final summary.

### 2.4 Reduction

The reduction process aims at reducing the original content to produce a summary containing simpler and more informative sentences. Sentence simplification removes extraneous information from a sentence, to ensure that it contains just as much information as needed. This process removes from each sentence specific structures, phrases or expressions explaining previously mentioned information. This feature determines that this information can be removed, allowing for novel one to be included in the summary.

The compression algorithm takes the summary candidate sentences. First, the compression rate (computed in words) is applied to that set of sentences, defining the first set of summary sentences. Afterwards, simplification rules are

applied to the set of summary sentences. Then, since the simplification process removes words from sentences, more sentences are added to the set of summary sentences, until the compression rate is once again attained. This procedure is repeated while the maximum number of words determined by the compression rate has not been achieved.

### 2.5 Presentation

The summary is delivered to the user in the form of a text file.

## 3 Outline

An outline of the script to be followed is presented here.

**Step 1** : Presentation of the set of documents serving as example.
**Step 2** : Presentation of an ideal summary for that set of documents.
**Step 3** : Execution of SIMBA over the example, and presentation of each stage of the summarization process, by demonstrating:
- **Step 3.1** : The sentences retrieved after the annotation process.
- **Step 3.2** : The sentences updated after the main scores computation.
- **Step 3.3** : The clusters resulting from the similarity clustering.
- **Step 3.4** : The sentences that have been discarded and the ones that will remain in the summarization process.
- **Step 3.5** : The global keywords retrieved from the set of sentences.
- **Step 3.6** : The clusters resulting from the keywords clustering.
- **Step 3.7** : The sentences that have been discarded and the ones that will remain in the summarization process.
- **Step 3.8** : The sentences filtered by length.
- **Step 3.9** : The compression algorithm, by demonstrating:
  - **Step 3.9.1** : The sentences to be simplified and their removable phrases.
  - **Step 3.9.2** : The sentences after being simplified.
  - **Step 3.9.3** : The sentences added after the simplification process.
  - **Step 3.9.4** : The repetition of the algorithm, while the compression rate has not been achieved.
- **Step 3.10** : Presentation of the final summary retrieved to the user.

**Step 4** : Evaluation of the summary produced by SIMBA when compared to the ideal summary for the example set.

## References


1. Branco, A. and Silva, J.: A Suite of Shallow Processing Tools for Portuguese: LX-Suite. Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL'06) (2006).
2. Silva, J., Branco, A., Castro, S., Reis, R.: Out-of-the-Box Robust Parsing of Portuguese. Proceedings of the 9th Encontro para o Processamento Computacional da Língua Portuguesa Escrita e Falada (PROPOR), 75–85, (2010).
3. Silveira, S. B., Branco, A.: Extracting Multi-document Summaries with a Double Clustering Approach. Proceedings of the 17th International Conference on Applications of Natural Language Processing to Information Systems, (to appear).