Rearrangement and Creation of New Corpora for Update and Compressive Summarization Tasks for Portuguese Language

Fernando Antônio Asevedo Nóbrega, Thiago Alexandre Salgueiro Pardo

Núcleo Interinstitucional de Linguística Computacional (NILC)
Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo
São Carlos, SP, Brazil

{fasevedo,taspardo}@icmc.usp.br

Abstract. *We introduce in this paper two corpora for new summarization tasks for Portuguese language. We report (i) the rearrangement of an existent multi-document summarization corpus in order to be used for update summarization and (ii) the creation of a corpus for training compressive summarization methods.*

1. Introduction

Automatic Summarization (AS) aims to automatically produce a condensed version of one or more related source texts/documents (Mani, 2001). In the beginning, the main goal of AS was the production of a summary for only one text, in a task named single document summarization. However, nowadays, the multi-document summarization, which aims to produce a unique summary from a cluster of texts on the same topic, has received more attention, as it is more common in the online scenery, in which the reader/user frequently finds many related content.

More recently, new summarization tasks have arisen. In the Document Understanding Conference (DUC) of 2007[1], the update summarization task was introduced. This task aims to produce summaries under the assumption that the reader has already read some previous related texts: the summary should contain only the more relevant and updated information. In the DUC 2007 dataset for update summarization, there were three text clusters/collections (named A, B and C) sorted in chronological order. An update summary for each cluster should be produced under the assumption that the reader had already read the previous ones (Witte et al., 2007). The update summarization task was also present in the Text Analysis Conferences (TAC) since 2008, but only two text clusters (A and B) were adopted.

Independently of the kind of the AS task, we also may define the methods based on their summary synthesis approach, as extractive, abstractive or, yet, compressive approaches. In the first one, the methods pick some sentences from the source texts and organize them in the output, so that there is not new content production. In the abstractive approach, the systems produce new sentences for the output summary based on the source texts, using rewriting operations. Finally, in the compressive one, the methods pick some sentences and eventually compress/reduce some of them before adding them to the summary, being an intermediate step between extractive and abstractive approaches. As the extractive approaches may produce summaries with some cohesive problems and redundant or irrelevant segments and the abstractive

[1] `http://www.nist.gov/guidelines/2007.html`

methods still demand more efforts in order to scale up and to produce summaries with satisfactory quality, the compressive summarization may be a good path to follow, as some previous researches have shown (Jing, 2000; Kawamoto and Pardo, 2010; Berg-Kirkpatrick et al., 2011; Li et al., 2013; Almeida and Martins, 2013).

For the Portuguese language, there are many AS investigations based on the extractive approaches for single and multi-document tasks, as (Pardo, 2002; Rino et al., 2004; Muller et al., 94; Leite et al., 2007; Antiqueira et al., 2009; Castro Jorge and Pardo, 2011; Ribaldo et al., 2012; Silveira and Branco, 2012; Nóbrega et al., 2014; Cardoso and Pardo, 2015; Ângelo Abrantes Costa and Martins, 2015; Cardoso and Pardo, 2016; Ribaldo et al., 2016). Many of them have used the CSTNews corpus (Aleixo and Pardo, 2008; Cardoso et al., 2011), in which there are human summaries based on the extractive and abstractive approaches. However, this corpus is not prepared for the update summarization task. For compressive summarization, there is the Priberam Compressive Summarization Corpus (PCSC) (Almeida et al., 2014), with human compressive summaries. However, PCSC is still relatively small in order to train methods for sentence compression, which is an important step for compressive-based AS methods.

In this paper, we introduce two different corpora in order to assist the extension and deepening of AS investigation in Portuguese for different tasks, as the update summarization and the compressive approach. The first one, named CSTNews-Update, is a rearrangement/adaptation of the CSTNews corpus for subsidizing research on update summarization. The second one is a dataset with pairs of original (long) sentences and their reduced versions.

We introduce the CSTNews-Update corpus in Section 2 and the corpus for sentence compression in Section 3. We present some final remarks in Section 4.

2. CSTNews-Update

CSTNews-Update is a different setup of the CSTNews corpus, which has been used in many researches of single and multi-document summarization methods for the Portuguese language. It aims to allow the investigation of update summarization methods in this language.

In similar way to the datasets that were used in the DUC/TAC conferences, each text set in CSTNews-Update has a text collection A and another one B, so that we envision the production of an update summary from the B cluster under the assumption that the reader has already read the texts in A. The maximal amount of texts in each cluster in CSTNews-Update is three, and the minimal is one or two for the A and B clusters, respectively. This way, all the update summaries that may be produced for our corpus are also multi-document summaries. We follow the DUC/TAC strategy, in which there are always more than one text in the recent cluster.

Based on the update summarization definition and in the above restrictions, we have defined 58 distinct text sets for the CSTNews-Update by using two different approaches: intra-cluster and inter-cluster. In the first case, we have picked clusters with three texts (in a total of

39) from CSTNews, so that for each set we have labeled the oldest text as the collection A and the others as the collection B. In the second way, we manually grouped pairs of different clusters from CSTNews with similar subjects (in a total of 19), in which, for each set, the cluster with the oldest texts was considered the collection A. In total, CSTNews-Update has 3.320 sentences, 49.449 words and 225 texts (95 that were labeled as A and 130 as B texts).

All the texts were sorted by the timestamp they were published. For some cases, in which this information was not available, we have manually sorted the documents by analyzing the temporal details of the texts. An interesting feature of CSTNews-Update are the differences in the timestamp distances (from seconds to days) among old and new documents in their clusters. We believe that this feature simulates cases of real world, in which the users may read sequential texts that have low timestamp difference and also read others that have larger differences. As expected, the timestamp differences among documents are low in the intra-cluster approach and larger in the inter-cluster. The maximal difference is approximately 9 days and the average is $\approx$ 175.51 hours. This way, CSTNews-Update corpus also enables investigations about the impact of the published time of documents to find updated and new information.

It is also important to say that, once we used the CSTNews as the basis for our dataset, most of the linguistic knowledge that was manually identified in this corpus is also available in CSTNews-Update, as follows: nouns and verbs that were disambiguated (Nóbrega and Pardo, 2012; Sobrevilla Cabezudo et al., 2014) with synsets of the Princeton WordNet (Fellbaum, 1998); discursive relations based on the Rhetorical Structure Theory (Mann and Thompson, 1987); subtopic segmentation (Cardoso et al., 2013); and informative aspects based on the guided summarization task that was introduced in the TAC of 2011 (Owczarzak and Dang, 2011); among other annotations. There are also discursive relations from the Cross-document Structure Theory (Radev, 2000). However, some of these are missed in the new corpus, once these relations occur among sentences from different texts from a same collection. Thus, in our datasets, we just have CST relations in the inter-clusters.

3. A Corpus for Sentence Compression

As we have said before, although the PCSC corpus has human summaries that were made based on the compressive approach, the amount of data is not enough for training sentence compression methods. Filippova et al. (2015), for example, used a dataset with 12.000 pairs of long sentences and their respective reduced versions, while PCSC has approximately 900 pairs, as collected by (Nóbrega and Pardo, 2016). This way, based on the procedure that was proposed by (Filippova and Altun, 2013), we analyzed the titles and first sentences of 1.008.356[2] documents that were automatically collected from the G1[3] news portal and selected the appropriate pairs.

[2] This number is the amount of documents that were collected, filtered and processed until this paper.

[3] http://www.g1.com.br

Accepted pair
Padre lança livros sobre teatro e centro histórico de Santarém.
O padre Sidney Canto lança, na sexta-feira (17) , livros sobre a o teatro e centro histórico de Santarém , oeste do Pará.
Unaccepted pair
Rodoviária de Porto Alegre tem 700 ônibus extras para o Carnaval.
Para atender a demanda no feriadão de Carnaval , a Rodoviária de Porto Alegre conta com 700 õnibus extras.

Figure 1. Example of candidate pairs of long and reduced sentences

Here, we just consider those candidates in which the title may be produced by a sequence of token deletions from the first sentence in its respective document. Changing the position of one or more words was not accepted. In the Figure 1, we may see candidate pairs that were accepted or that were not.

Even the definition of SC above have been used for many authors in this field, as (Turner and Charniak, 2005; Martins and Smith, 2009; Kawamoto and Pardo, 2010; Almeida and Martins, 2013; Thadani and McKeown, 2013; Cordeiro et al., 2013; Filippova et al., 2015; Nóbrega and Pardo, 2016), this procedure is a simplification of the Compression made by humans, in which more sophisticated steps may be applied, as the use of new words, synonyms or syntactical transformations. However, it is important to say this simplification is required as the initial stage for the production of SC systems.

In the initial processing of the candidate pairs (the respective title and first sentence of each document), we have found some very similar sentential pairs that were not considered appropriate because of some very tiny differences. Thus, we have applied some rules based on frequent patterns that occur in our dataset in order to normalize the candidates, as follows:

- A period (.) was added in the end of all the sentences that did not end with some punctuation mark.
- The comma (,) was used as decimal separator for the float numbers (it is the pattern used in Portuguese).
- Blank spaces were added around all the punctuation marks in order to avoid mistakes based on some wrong tokenization of the pairs (e.g., “),” was changed to “) , ”).
- All the quoted texts were normalized to use simple quote (’).
- All the acronyms that did not occur between parentheses were normalized to the parenthetical form in the sentences in which they occured in different ways only (as “(SP)” and “–SP–”).

After the above process, we found 7,056 pairs of original sentences and the respective reduced versions. However, we removed 32 of them because we considered as irrelevant samples for the sentence compression task. We removed: candidate pairs in which the titles were just names of people, very short candidates (5 tokens), sentences that began with some newspaper mark (update, correct link, infograph, etc.), and sentences in which the content was organized as a list of items or similar to tables (currency values, taxes, etc.).

Figure 2 shows a histogram with the sizes of the sentences in our dataset. As we may see, most of the sentences has more than 20 and less than 60 tokens. The average compression rate of the reduced sentences is 57% in relation to the original sentences (in number of words).

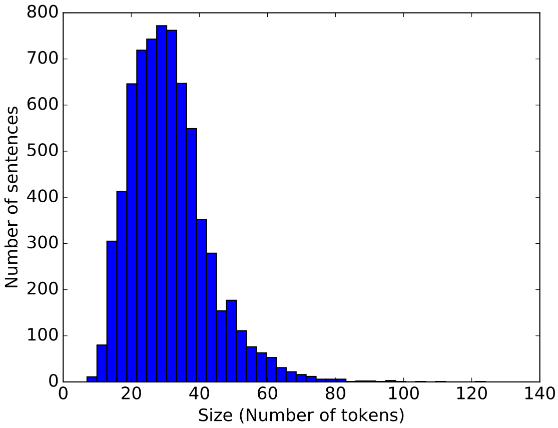

Figure 2. Distribution of tokens per sentence in the dataset

4. Final Remarks

Table 1 synthesizes the details of the two corpora that we have introduced in this paper: the CSTNews-Update[4] and the sentence compression corpus[5]. Although the adopted strategies for building these corpora are not standard ones, they showed to be valuable and with a good cost/benefit ratio.

We expect that these corpora allow more investigations in the AS area, fostering the development of interesting and relevant applications for the Portuguese language.

Table 1. Synthetic view of the corpora

	CSTNews-Update	Sentence compression corpus
Number of Clusters	58	–
Number of Documents	225	–
Number of Sentences	3,320	7,024 (long and reduced sentences)
Number of Tokens	49,449	218,789

5. Acknowledgments

The authors are grateful to CAPES and FAPESP for supporting this work.

[4] `https://github.com/fernandoasevedo/CSTNews-Update`

[5] `https://github.com/fernandoasevedo/Sentence-Compression-for-Portuguese-Language`

References

Aleixo, P. and T. A. S. Pardo (2008). *CSTNews: Um Córpus de Textos Jornalísticos Anotados segundo a Teoria Discursiva Multidocumento CST (Cross-document Structure Theory).* Technical Report 326, Instituto de Ciências Matemáticas e de Computação.

Almeida, M. B., M. S. C. Almeida, A. F. T. M. H. Figueira, P. Mendes, and C. Pinto (2014). A new multi-document summarization corpus for european portuguese. In *Language Resources and Evaluation Conference (LREC'14)*, Reykjavik, Iceland, pp. 1–7.

Almeida, M. B. and A. F. T. Martins (2013). Fast and robust compressive summarization with dual decomposition and multi-task learning. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, Sofia, Bulgaria, pp. 196–206. Association for Computational Linguistics.

Ângelo Abrantes Costa, M. and B. Martins (2015). Uma comparação sistemática de diferentes abordagens para a sumarização automática extrativa de textos em português. *Linguamática* Vol. *7* N. 1, pp. 23–40.

Antiqueira, L., O. N. Oliveira Jr., L. d. F. Costa, and M. d. G. V. Nunes (2009). A complex network approach to text summarization. *Information Sciences* Vol. *179* N. 5, pp. 584–599.

Berg-Kirkpatrick, T., D. Gillick, and D. Klein (2011). Jointly learning to extract and compress. In *Proceedings of the International Conference on Computational Linguistics*, Portland, Oregon, pp. 481–490. Association for Computational Linguistics.

Cardoso, P. and T. A. S. Pardo (2015). Joint semantic discourse models for automatic multi-document summarization. In *Proceedings of the 10th Brazilian Symposium in Information and Human Language Technology (STIL)*, Natal, RN, Brazil, pp. 81–90.

Cardoso, P. and T. A. S. Pardo (2016). Multi-document summarization using semantic discourse models. *Processamiento de Lenguaje Natural Vol. 56*, pp. 57–64.

Cardoso, P., M. Taboada, and T. Pardo (2013). Subtopic annotation in a corpus of news texts – steps towards automaticsubtopic segmentation. In *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology (STIL)*, Cuiabá, MT, Brazil, pp. 49–58.

Cardoso, P. C. F., E. G. Maziero, M. L. R. Castro Jorge, E. M. R. Seno, A. Di Felippo, L. H. M. Rino, M. d. G. V. Nunes, and T. A. S. Pardo (2011). CSTNews – a discourse-annotated corpus for single and multi-document summarization of news texts in brazilian portuguese. In *Anais do III Workshop "A RST e os Estudos do Texto"*, Cuiabá, MT, Brasil, pp. 88–105. Sociedade Brasileira de Computação.

Castro Jorge, M. L. and T. A. S. Pardo (2011). A generative approach for multi-document summarization using the noisy channel model. In *Proceedings of the 3rd RST Brazilian Meeting*, Cuiabá/MT, Brazil, pp. 75–87.

Cordeiro, J., G. Dias, and P. Brazdil (2013). *Rule Induction for Sentence Reduction*, pp. 528–539. Berlin, Heidelberg: Springer Berlin Heidelberg.

Fellbaum, C. (1998). *WordNet An Eletronic Lexical Database.* MIT Press.

Filippova, K., E. Alfonseca, C. Colmenares, L. Kaiser, and O. Vinyals (2015). Sentence compression by deletion with LSTMs. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Lisbon, Portugal, pp. 360–368. Association for Computational Linguistics.

Filippova, K. and Y. Altun (2013). Overcoming the lack of parallel data in sentence compression. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Seattle, Washington, USA, pp. 1481–1491. Association for Computational Linguistics.

Jing, H. (2000). Sentence reduction for automatic text summarization. In *Proceedings of the Sixth Conference on Applied Natural Language Processing*, ANLC '00, Seattle, Washington, pp. 310–315. Association for Computational Linguistics.

Kawamoto, D. and T. A. S. Pardo (2010). Learning sentence reduction rules for brazilian portuguese. In *Proceedings of the 7th International Workshop on Natural Language Processing and Cognitive Science*, Funchal, Madeira, Portugal, pp. 90–99.

Leite, D. S., L. H. M. Rino, T. A. S. Pardo, and M. das Graças V. Nunes (2007). Extractive automatic summarization: Does more linguistic knowledge make a difference? In *Proceedings of the HLT/NAACL Workshop on TextGraphs-2: Graph-BasedAlgorithms for Natural Language Processing*, Rochester, NY, USA, pp. 17–24. Association for Computational Linguistics.

Li, C., F. Liu, F. Weng, and Y. Liu (2013). Document summarization via guided sentence compression. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Seattle, Washington, USA, pp. 490–500. Association for Computational Linguistics.

Mani, I. (2001). *Automatic Summarization*, Vol. 3. John Benjamins Publishing Company.

Mann, W. C. and S. A. Thompson (1987). Rhetorical structure theory: A theory of text organization. Technical report, ISI/RS-87-190.

Martins, A. F. T. and N. A. Smith (2009). Summarization with a joint model for sentence extraction and compression. In *Proceedings of the NAACL HLT Workshop on Integer Linear Programming for Natural Language Processing*, pp. 1–9.

Muller, E., J. Granatyr, and O. Lessing (75–94). Comparativo entre o algoritmo de luhn e o algoritmo gistsumm para sumarização de documentos. *Revista de Informática Teórica e Aplicada* Vol. *22* N. 1, pp. 584–599.

Nóbrega, F. A. A., V. Agostini, R. T. Camargo, A. Di Felippo, and T. A. S. Pardo (2014). Alignment-based sentence position policy in a news corpus for multi-document summarization. In *Proceedings of the 11st International Conference on Computational Processing of Portuguese (LNAI 8775)*, São Carlos, SP, Brazil, pp. 6–9.

Nóbrega, F. A. A. and T. A. Pardo (2012). Explorando métodos de uso geral para desambiguação lexical de sentidos para a língua portuguesa. In *Anais do 9o Encontro Nacional de Inteligência Artificial – ENIA*, Curitiba, PR, Brazil, pp. 1–12.

Nóbrega, F. A. A. and T. A. S. Pardo (2016, jul). Investigating machine learning approaches for sentence compression in different application contexts for portuguese. In *Proceedings* *of the XII International Conference on the Computational Processing of Portuguese*, Tomar, Portugal, pp. 245–250.

Owczarzak, K. and H. Dang (2011). Overview of the TAC 2011 summarization track: Guided task and AESOP task. In *Proceedings of the Text Analysis Conference*, NIST. National Institute of Standards and Technology.

Pardo, T. A. S. (2002). DMSumm: Um gerador automático de sumários. Master's thesis, Universidade Federal de São Carlos.

Radev, D. R. (2000). A common theory of information fusion from multiple text sources step one: Cross-document structure. In *Proceedings of 1st ACL SIGDIAL Workshop on Discourse and Dialogue*, Hong Kong, China, pp. 10. Association for Computational Linguistics.

Ribaldo, R., A. T. Akabane, L. H. M. Rino, and T. A. S. Pardo (2012). Graph–based methods for multi-document summarization: Exploring relationship maps, complex networks and discourse information. In *Proceedings of the 10th International Conference on Computational Processing of Portuguese (LNAI 7243)*, Coimbra, Portugal, pp. 260–271.

Ribaldo, R., P. F. Cardoso, and T. A. S. Pardo (2016). Exploring the subtopic-based relationship map strategy for multi-document summarization. *Journal of Theoretical and Applied Computing – RITA* Vol. *23* N. 1, pp. 183–211.

Rino, L. H. M., T. A. S. P. andCarlos Nascimento Silla Jr, C. A. A. Kaestner, and M. Pombo1 (2004). A comparison of automatic summarization systems for brazilian portuguese texts. In *Proceedings of the 17th Brazilian Symposium on Artificial Intelligence*, Sao Luis, Maranhao, Brazil, pp. 235–244. Springer Berlin Heidelberg.

Silveira, S. and A. H. Branco (2012). Enhancing multi-document summaries with sentence simplification. In *Proceedings of the 14th International Conference on Artificial Intelligence*, pp. 742–748.

Sobrevilla Cabezudo, M. A., E. G. Maziero, J. W. Souza, M. S. Dias, P. C. F. Cardoso, P. F. Balage, V. Agostini, F. A. A. N. C. D. Barros, A. Di Felippo, and T. A. SalgueiroPardo (2014). Anotação de sentidos de verbos em notícias jornalıisticasem português do brasil. In *Proceedings of the XII Encontro de Linguística de Corpus – ELC*, Uberlândia, MG, Brazil, pp. 1–7.

Thadani, K. and K. McKeown (2013). Sentence compression with joint structural inference. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pp. 65–74.

Turner, J. and E. Charniak (2005). Supervised and unsupervised learning for sentence compression. In *Proceedings of the 43rd Annual Meeting on Association for Computational*, pp. 290–297.

Witte, R., R. Krestel, and S. Bergler (2007). Generating update summaries for duc 2007. In *Proceedings of the Document Understanding Conference*, Rochester, New York USA, pp. 5.