Alignment-based Sentence Position Policy in a News Corpus for Multi-document Summarization

Fernando A. A. Nóbrega[1], Verônica Agostini[1], Renata T. Camargo[2], Ariani Di Felippo[2], Thiago A. S. Pardo[1]

Núcleo Interinstitucional de Linguística Computacional (NILC)
[1] Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo
[2] Departamento de Letras, Universidade Federal de São Carlos

{fasevedo, agostini}@icmc.usp.br, renatatironi@hotmail.com, arianidf@gmail.com, taspardo@icmc.usp.br

Abstract. This paper presents an empirical investigation of sentence position relevance in a corpus of news texts for generating abstractive multi-document summaries. Differently from previous work, we propose to use text-summary alignment information to compute sentence relevance.

1 Introduction

Multi-Document Summarization (MDS) is the task of automatically producing a unique summary from a group of source texts (documents) on the same topic [11][14]. It is a relatively new area (dating back to 1995 [13]) and brings old and well-known scientific challenges from the first studies in summarization in the 50s as well as introduces new and exciting challenges, e.g., to deal with redundant, complementary and contradictory information, to normalize different writing styles and referring expression, to balance different perspectives and sides of the same events and facts, to properly deal with evolving events and their narration in different moments, and to arrange information pieces from different texts to produce coherent and cohesive summaries, among others.

MDS, as many other Natural Language Processing tasks, may benefit from specialized corpora, as the ones built for the tasks of the Text Analysis Conferences. Such corpora usually contain large groups of source texts and human summaries, subsidizing researches on the nature and the phenomena that happen in summaries as well as allowing the development/training and comparative evaluation of state-of-the-art summarization systems.

In this paper, we report an empirical study of sentence position relevance for summarization, using a corpus of news texts and their abstractive multi-document summaries – the CSTNews corpus [1][5] – to learn summarization preferences, building on some previous work [10][9]. Giving one more step from where these works stopped, we use one of the corpus available annotations – the text-summary alignment information – to determine, in a more precise way, a robust sentence position policy to the selection of sentences that may compose the summaries. For

now, we are only worried on characterizing such policy, in a theoretical perspective. To the best of our knowledge, this is the first attempt carried out for a corpus in Brazilian Portuguese.

This paper is organized as follows. Section 2 introduces the CSTNews corpus and its annotation layers. Section 3 presents our study on sentence position and the achieved results.

2 The CSTNews Corpus

The CSTNews corpus [1][5] is a reference corpus for MDS composed of 50 clusters of news texts in Brazilian Portuguese (BP). Each cluster contains two or three source texts on the same topic, which were manually selected from on-line mainstream Brazilian news agencies as *Folha de São Paulo*, *Estadão*, *O Globo*, *Gazeta do Povo*, and *Jornal do Brasil*. Besides the original texts, each cluster conveys a manual (abstractive) single-document summary (with 30% compression rate) for each document in the cluster, a manual (abstractive) multi-document summary for the cluster and its corresponding manual extractive summary, and an automatic multi-document summary, produced by a state-of-the-art system for Portuguese [7].

The corpus also has annotated versions of the source texts and multi-document summaries in different linguistic levels, and according to different linguistic theories and models. Specifically, the source texts are manually annotated in different ways for discourse organization, following both the Rhetorical Structure Theory [12] and Cross-document Structure Theory [16]. They also have other manual annotations: their temporal expressions annotated and resolved, their most frequent nouns indexed to their corresponding senses in Princeton Wordnet, and subtopic segmentations and the keywords for each subtopic. Automatic annotations are also available, as the syntactical analyses for each sentence, produced by the PALAVRAS parser [3].

More recently, the corpus had the source text sentences aligned to the sentences of the manual multi-document summaries that shared some information with the formers. Therefore, since each summary comes from 1 or more texts, each sentence in the summaries might be aligned to more than 1 sentence in the texts. Most of the sentences in the summaries were aligned up to 5 sentences from the texts. In general, 42% of the sentences of the texts were aligned to some sentence in the summaries. As an example of alignment, in Table 1 we show a sentence in a summary that was aligned to 2 sentences from different source texts (translated from Portuguese):

Table 1. Example of alignment

Sentence from the summary	*Sentences from source texts*
Brazil will not be part of the torch relay, which includes 20 countries.	The torch will pass by twenty countries, but Brazil is not in the Olympic way. Brazil is not part of the path of the Olympic torch.

The alignment was performed by two computational linguistics, showing a 0.831 inter-annotator agreement kappa value [6], indicating that the annotation is reliable.

3 Sentence Position Policy

In [10] it was defined what was called "sentence position policy" for summary composition. The authors attributed a score for each sentence position in the texts and ranked the sentence relevance in terms of this score. Therefore, a good summary should be composed of the sentences from the best ranked positions. To compute the score of each sentence position, the authors counted the number of topic keywords in all the sentences in each specific position in a group of texts and averaged such number by the number of sentences in that position. They evaluated the resulting sentence position policy for single-document summarization and achieved good results. [9] produced more refined results by counting and averaging, for each sentence position, the number of Summary Content Units (SCUs) in the sentences. The SCUs were those available according to the pyramid method [15]. The authors evaluated the resulting policy for multi-document summarization and produced state of the art results. The above works demonstrated that position policies are worthy to pursue for corpus characterization and summarization.

In this paper, we build on the previous work by refining even more the calculation of the sentence position policy. We use the text-summary alignment information in the CSTNews corpus to better compute the sentence position rank for multi-document summarization. For each sentence position, we count and average the number of alignments they have with the corresponding manual multi-document summary.

Alignments may be more informative than topic keywords or SCUs for the envisioned task. While topic keywords are at the lexical level and SCUs are more conceptual, the alignments may represent any of this information. We consider two versions of the alignment-based policy: one counting the total number of alignments among the sentences and other counting only once the alignments for a pair of sentences, does not mattering how many alignments they have.

For comparison purposes, we also created an alternative, and simpler, position policy. As our corpus does not present topic keywords or pyramid SCUs, we used the own words (excluding stopwords and punctuation marks) in the manual multi-document summaries to score each sentence position. In this case, each sentence position is scored as the average of different words from the summary that it contains.

Figures 1 and 2 show the graphics for sentence position policy by counting summary words, using the absolute sentence positions and their normalized versions, respectively. The normalized graphic allows to make text sizes uniform, ranging from 0 to 1, and, therefore, resulting in fairer analyses. In each graphic, the blue line represents the average values for each sentence position, while the green and red lines incorporate the standard deviation above and below the blue line, respectively.

One may see from the graphics that the first sentences are more relevant than the others. There is also a variation among the positions 20 and 30, approximately, accompanied by a variation of the standard deviation too, showing that it is not possible to fully trust in such specific sentence positions to compose good summaries. It is also possible to realize that the normalized sentence positions show the same behavior of the non-normalized version.

Figures 3 and 4 show the graphics for the sentence position policy with the two versions of alignment information – considering the alignments only once for each sentence position and all the alignments for each sentence, respectively. In relation to

the graphics for the summary words, similar behavior may be observed for the alignments. However, one may notice that the curves fall softer, indicating that the first sentences of the texts have more relevance than the others, but are closer to one another than the word-based computation revealed. We do not show the normalized versions of the graphics because they also show similar behavior.

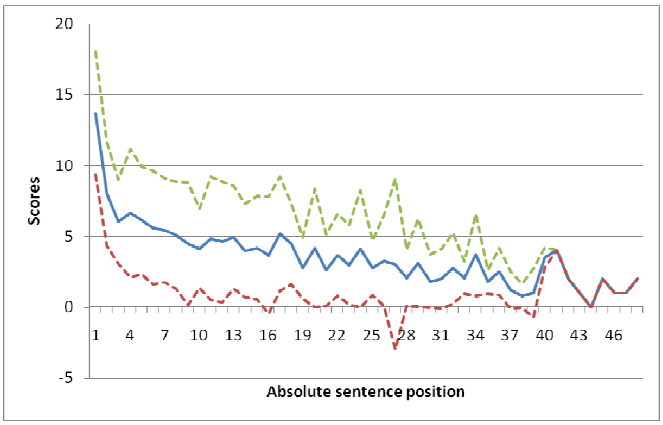

Fig. 1. Absolute sentence position policy with summary words

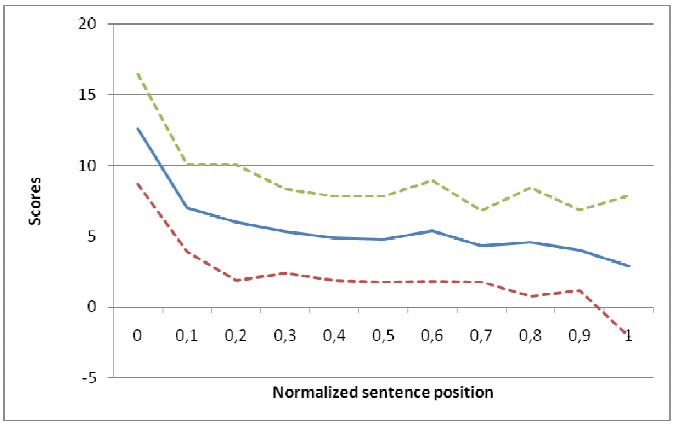

Fig. 2. Normalized sentence position policy with summary words

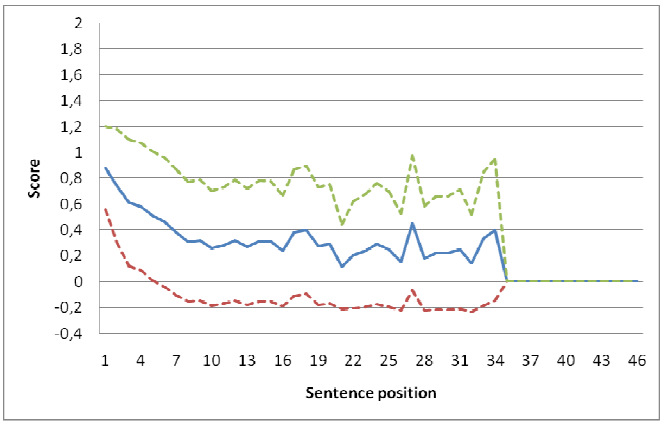

Fig. 3. Sentence position policy with alignments considered only once

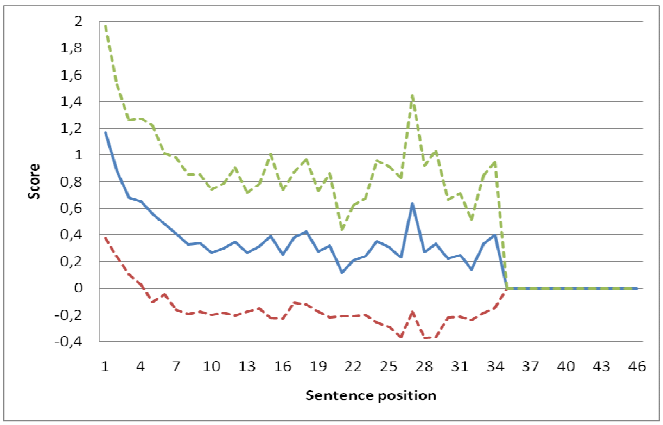

Fig. 4. Sentence position policy with all the alignments

From this study, as expected, one may conclude that, for the CSTNews corpus, the first sentences in the texts have more content that is expressed in the multi-document summaries. More than this, it is interesting to see that, from the 10th position on, it is more difficult to differentiate sentence relevance in terms of the position. In fact, after the most important initial sentences, there is almost a plateau of sentence relevance marked by disturbing areas (as standard deviation evidences).

The achieved results also evidence our previous findings for the CSTNews. In [4], it was verified that 89% of the first sentences in the source texts were aligned to the sentences in the summary. They are, therefore, very good candidate sentences to compose multi-document extractive summaries, as several works on summarization have showed (see, e.g., [2][8]) and [9] demonstrated to produce state-of-the-art results.

Future work includes deepening this study with other annotation layers available in the corpus (as the discourse annotations) and applying these strategies to produce automatic summarization systems, which would probably be strong baselines in the area.

Acknowledgments

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

References

[1] Aleixo, P. and Pardo, T.A.S. (2008). *CSTNews: Um Córpus de Textos Jornalísticos Anotados segundo a Teoria Discursiva Multidocumento CST (Cross-document Structure Theory)*. ICMC-USP Technical Report N. 326, 12p.

[2] Baxendale, P.B. (1958). Machine-made index for technical literature – an experiment. *IBM Journal*, pp. 354-361.

[3] Bick, E. (2000). *The Parsing System Palavras - Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Famework*. PhD Thesis, Aarhus University Press.

[4] Camargo, R.T. (2013). *Investigação de Estratégias de Sumarização Humana Multidocumento*. MSc Dissertation. Departamento de Letras, Universidade Federal de São Carlos. 133p.

[5] Cardoso, P.C.F.; Maziero, E.G.; Jorge, M.L.C.; Seno, E.M.R.; Di Felippo, A.; Rino, L.H.M.; Nunes, M.G.V.; Pardo, T.A.S. (2011). CSTNews - A Discourse-Annotated Corpus for Single and Multi-Document Summarization of News Texts in Brazilian Portuguese. In the *Proceedings of the 3rd RST Brazilian Meeting*, pp. 88-105.

[6] Carletta, J. (1996). Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics*, Vol. 22, N. 2, pp. 249-254.

[7] Castro Jorge, M.L.R. and Pardo, T.A.S. (2010). Experiments with CST-based Multidocument Summarization. In the *Proceedings of the ACL Workshop TextGraphs-5: Graph-based Methods for Natural Language Processing*, pp. 74-82.

[8] Edmundson, H.P. (1969). New methods in automatic extracting. *Journal of the ACM*, Vol. 16, N. 2, pp. 264-285.

[9] Katragadda, R.; Pingali, P.; Varma, V. (2009). Sentence Position revisited: A robust light-weight Update Summarization 'baseline' Algorithm. In the *Proceedings of the Third International Cross Lingual Information Access Workshop*, pp. 46-52.

[10] Lin, C.Y. and Hovy, E. (1997). Identifying Topics by Position. In the *Proceedings of the Fifth Conference on Applied Natural Language Processing*, pp. 283-290.

[11] Mani, I. (2001). *Automatic Summarization*. John Benjamins Publishing Co., Amsterdam.

[12] Mann, W.C. and Thompson, S.A. (1987). *Rhetorical Structure Theory: A Framework for the Analysis of Texts*. ISI Reprint Series ISI/RS-87-190, Information Sciences Institute.

[13] McKeown, K. and Radev, D.R. (1995). Generating summaries of multiple news articles.In the *Proceedings of the 18th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pp. 74-82.

[14] Nenkova, A. and McKeown, K. (2011). *Automatic Summarization*. Foundations and Trends in Information Retrieval Series. Now Publishers Inc.

[15] Nenkova, A.; Passonneau, R.; McKeown, K. (2007). The pyramid method: Incorporating human content selection variation in summarization evaluation. *ACM Transactions on Speech and Language Processing*, Vol. 4, N. 2, pp. 1-23.

[16] Radev, D.R. (2000). A common theory of information fusion from multiple text sources, step one: Cross-document structure. In the *Proceedings of 1st ACL SIGDIAL Workshop on Discourse and Dialogue*, pp. 74-83.