CORPUS ANALYSIS OF ASPECTS IN MULTI-DOCUMENT SUMMARIES – THE CASE OF NEWS TEXTS FROM "WORLD" SECTION

Renata T. Camargo[1], Erick G. Maziero[2], Thiago A. S. Pardo[2]
Núcleo Interinstitucional de Linguística Computacional (NILC)
[1] Departamento de Letras, Universidade Federal de São Carlos
[2] Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo
renatatironi@hotmail.com, {erickgm, taspardo}@icmc.usp.br

Introduction

Multi-Document Summarization (MDS) is the task of producing a unique summary for a group of texts on the same topic (Mani 2001). With the increasing amount of information, mainly in the web, summarization has become a very relevant task, fostering several researches that have produced many tools and resources, including corpora of texts and summaries.

Although researches in MDS date back to the 90s, there are few studies on the composition of multi-document summaries. However, it is well known that humans usually produce different summaries for the same group of texts, selecting varied information to include in the summaries according to what they judge to be more important. According to Owczarzak and Dang (2011), information importance is a subjective criterion and it is necessary to define general guidelines for summary production in order to have better human agreement on what should be in a summary. Based on corpus analysis, Owczarzak and Dang recommend, for instance, that summaries from texts of the "accidents and natural disasters" category should include information on the following aspects: what happened, when it happened, why it happened, who was affected, damages and countermeasures that were taken. They also make aspect recommendations for the "attacks", "health and safety", "endangered resources", and "trials and investigations" categories. A few other works also tried to study summary aspects. White et al. (2001) propose "aspect templates" for natural disasters too. Zhou et al. (2005) study the aspects that should appear in biographical summaries. Li et al. (2010) explore usual aspects in entity summaries in Wikipedia. In

information extraction-based MDS approaches, some works explicitly model the aspects to produce summaries (e.g., Radev and Mckeown 1998; White et al. 2001).

In this paper, we conduct a corpus analysis of manual multi-document summaries in the CSTNews corpus (Aleixo and Pardo 2008; Cardoso et al. 2011), which is a corpus of texts and summaries for Brazilian Portuguese. The corpus has 50 clusters of news texts from varied on-line news agencies (*Folha de São Paulo*, *Estadão*, *Jornal do Brasil*, *O Globo*, and *Gazeta do Povo*) from different sections (World, Politics, Science, Sports, and Daily News). Each cluster has from 2 to 3 texts on the same topic and presents manually built multi-document summaries (abstracts). In particular, we focus our analysis on summaries from the "world" section, trying to identify categories in this section and the recurrent aspects and their organization in summaries. We expect that such analysis may contribute to the linguistic characterization of manual summaries and subsidize future work on MDS.

The aspect/discourse organization of other sections in the same corpus – CSTNews – has been reported by other authors. Castro Jorge et al. (2012) study the "sports" section; Rassi et al. (2012) deal with texts from the "politics" section; Zacarias et al. (2012) handle the "daily news" section. All of these initiatives, including the work reported in this paper, were developed in the context of the SUCINTO[1] project (*summarization for clever information access*), which aims at investigating and exploring multi-document summarization strategies for providing a more feasible and intelligent access to on-line information.

The remaining of this paper is organized as follows: we report our corpus analysis and results in the next section; some final remarks are presented in the last section of this paper.

Corpus analysis

The analysis of the corpus was based on its annotation, which was performed by three human annotators, with experience in computational linguistics. All the summaries in the "world" section of the CSTNews corpus were annotated. The corpus contains 14 "world" clusters and, therefore, 14 summaries of this type. In average, each summary

[1] www.icmc.usp.br/~taspardo/sucinto

has 129 words. Figure 1 shows the categories that occur in the "world" clusters and their frequency. One may see that "world" includes accidents, attacks, legal and political decisions, and natural disasters. Natural disaster is the most frequent category, with 5 clusters.

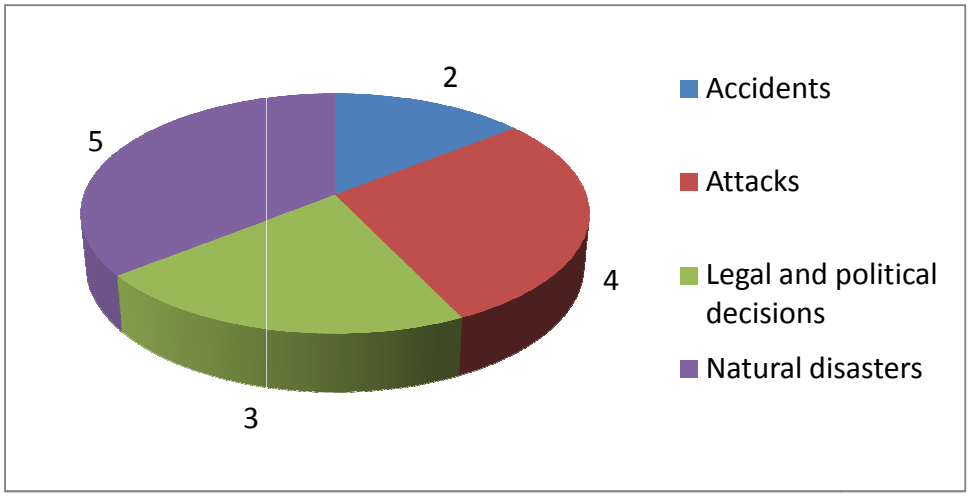

Fig. 1: Distribution of categories in the CSTNews "world" section

The annotation of each summary was performed by the three annotators together. It started with the same aspects proposed by Owczarzak and Dang (2011) for the summaries about accidents and natural disasters. It soon showed that some more aspects were necessary, namely: *history* about the fact/event being presented in the summary, *prediction* of some fact/event that will probably/possibly happen, *source* (person, news agency, etc.) of some information in the summary, and *perpetrator* (agent) of some action/event. When other related information (but not the main one focused in the summary) appeared, an additional "extra" mark was concatenated to the aspect to indicate it. As illustration, Figure 2 shows an annotated summary (translated from Portuguese and the original one). The aspects are shown in capital letters after the text passages they refer to, which are delimited by brackets.

Summary in English
[17 people died]WHO_AFFECTED after [an airplane crash]WHAT [in Democratic Republic of Congo.]WHERE [14 of these victims were passengers and three of them were crew members, all of them were Russian.]WHO_AFFECTED [Nobody survived.]DAMAGES_EXTRA
[The plane took off in Lugushwa and it was expected to land in Bukavu, but it fell down over a forest]WHERE [after colliding with a mountain because of bad weather.]WHY
[The plane was also carrying cargos and minerals.]WHAT_EXTRA

Summary in Portuguese (original)
[17 pessoas morreram]WHO_AFFECTED [após a queda de um avião]WHAT [na República Democrática do Congo.]WHERE [14 dessas vítimas eram passageiros e três membros da tripulação, todos de nacionalidade russa.]WHO_AFFECTED [Nenhuma vítima sobreviveu.]DAMAGES_EXTRA
[O avião saiu de Lugushwa a Bukavu e caiu sobre uma floresta]WHERE [após se chocar com uma montanha, prejudicado pelo mau tempo.]WHY
[O avião também levava cargas e minerais.]WHAT_EXTRA

Figure 2. Example of annotated summary

It is possible to see that the main information is the airplane crash (which is indicated by the *what* aspect), but there is also a *what_extra* aspect in the last sentence, since it is additional information (that the plane was carrying cargo and minerals).

Figure 3 shows the overall frequency of the used aspects (in alphabetical order) in the "world" summaries. As expected, the *what* aspect was the most frequent one, occurring 26 times. Figures 4 to 7 show the same distribution by category.

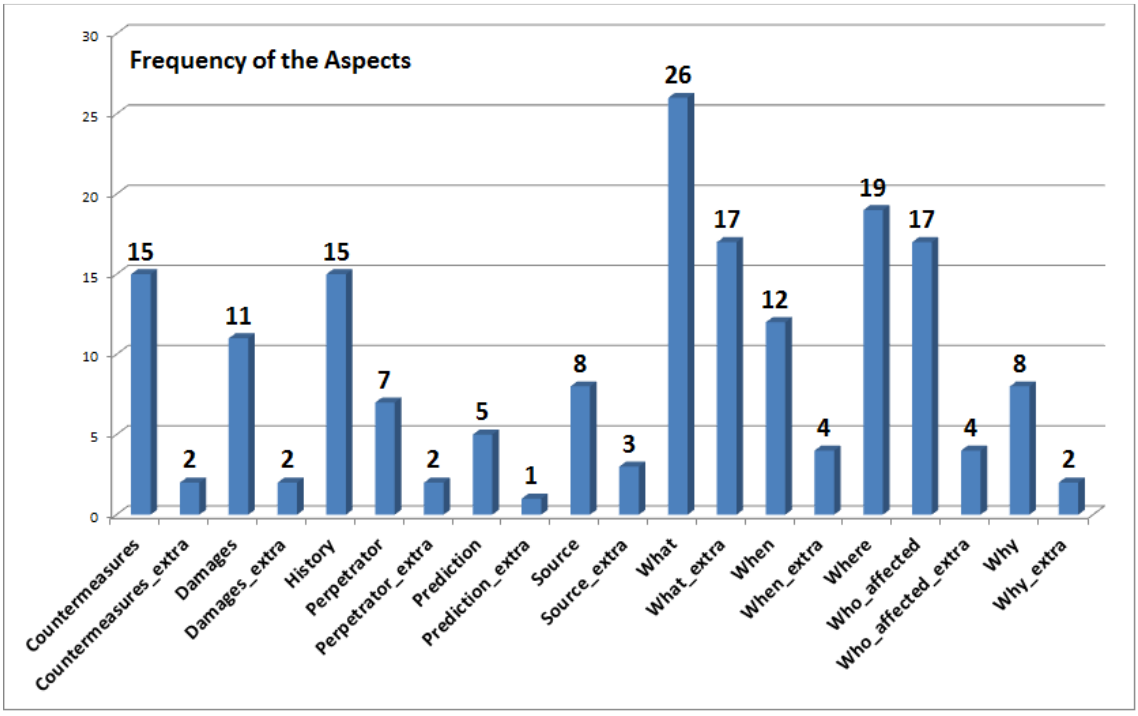

Fig. 3: Frequency of aspects in the “world” summaries

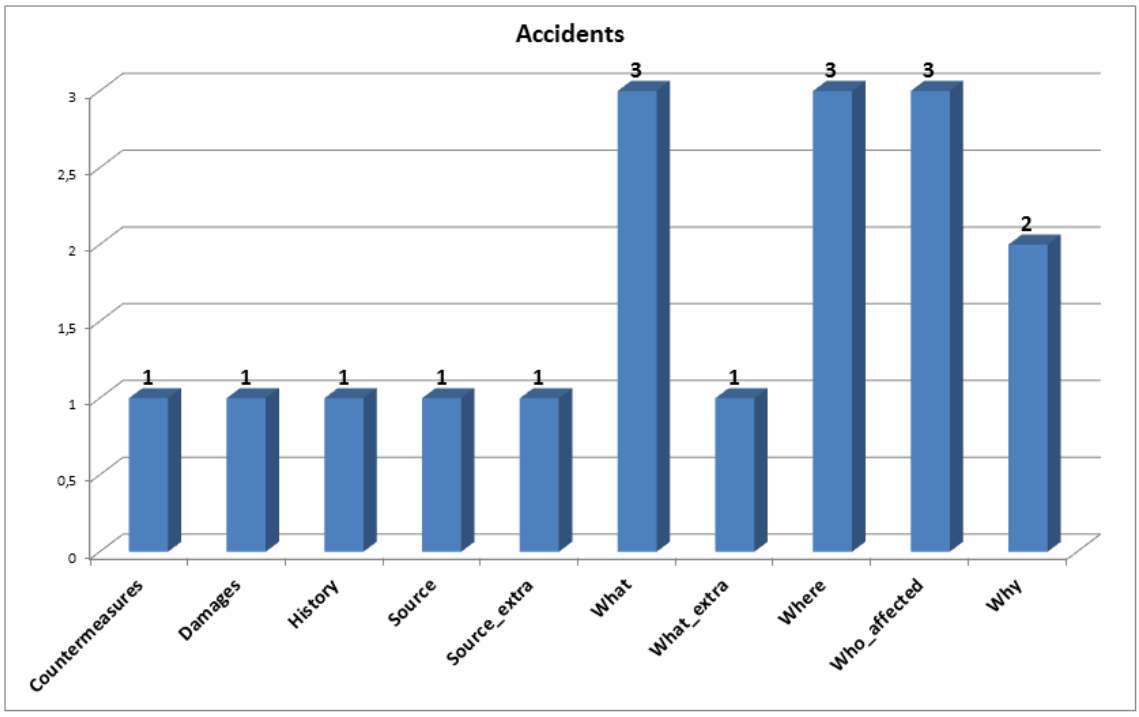

Fig. 4: Frequency of aspects in “Accidents” category

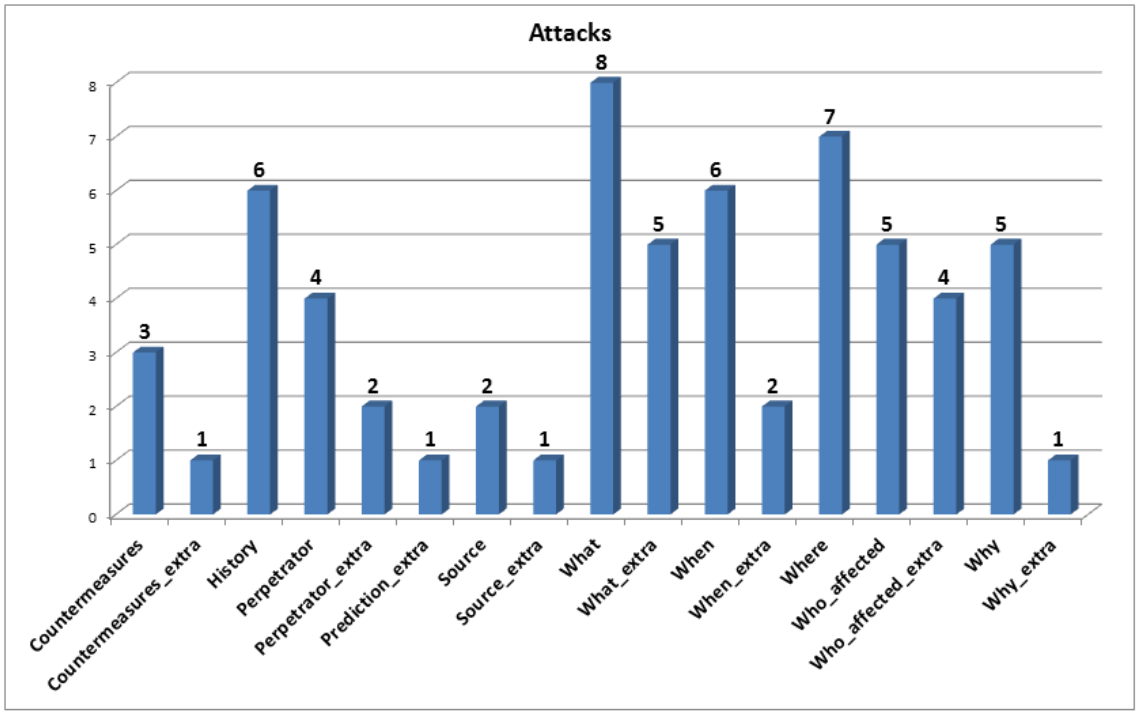

Fig. 5: Frequency of aspects in “Attacks” category

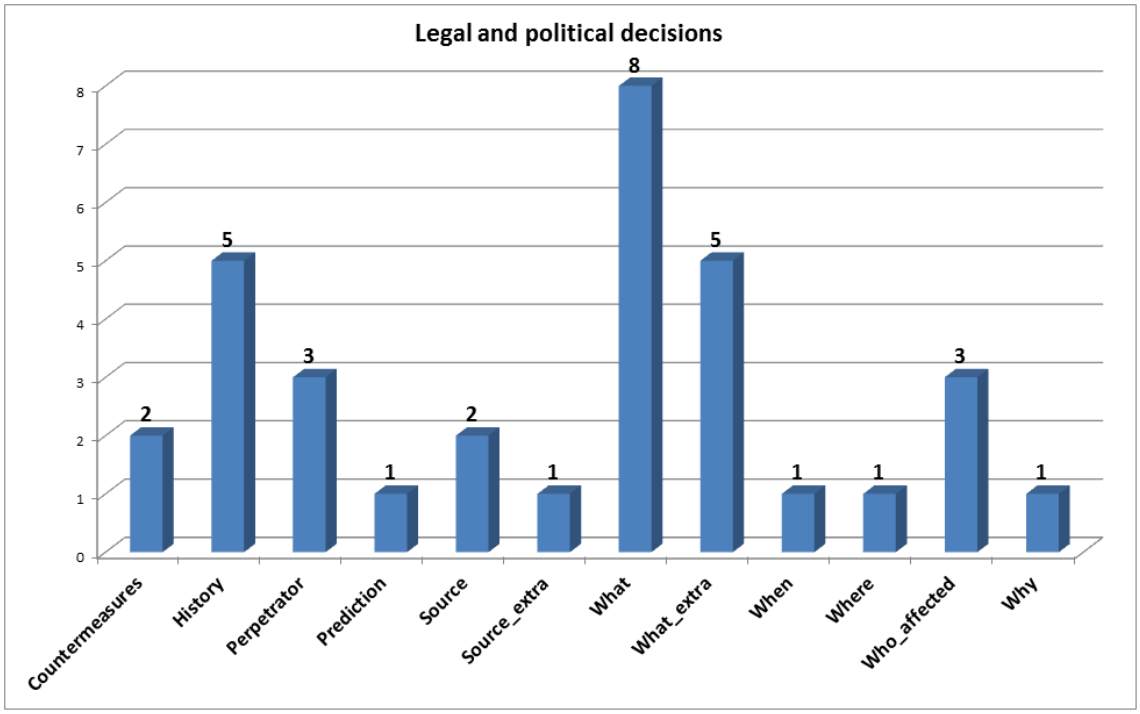

Fig. 6: Frequency of aspects in “Legal and political decisions” category

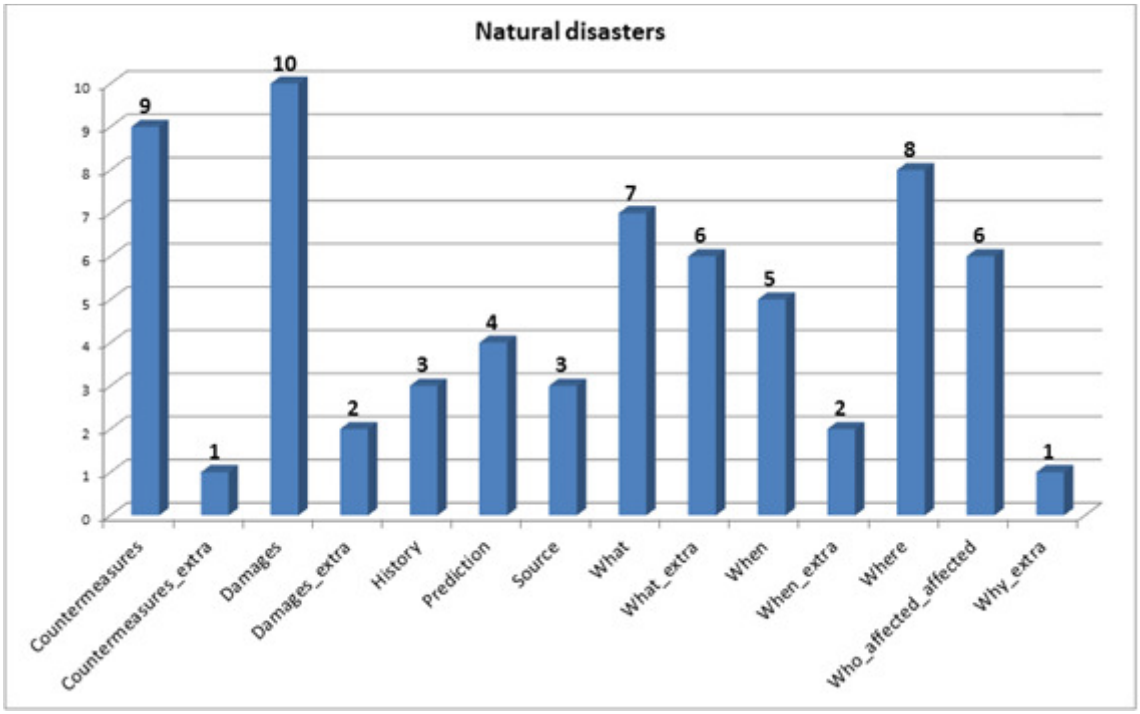

Fig. 7: Frequency of aspects in "Natural disasters" category

Analyzing the annotated corpus and the graphics before, it is possible to conclude that some aspects are more usual than others and that there are partial orderings among some of them. Our conclusions are synthesized in Table 1.

Table 1: Aspect analysis in the summaries

	Accidents	**Attacks**	**Legal and Political Decisions**	**Natural Disasters**
For all summaries				
In common	What, Where, Who_affected, Why	What, Where, Who_affected	What, Perpetrator	What, Where, Who_affected, Countermeasures, Damages
In the 1st paragraph	What, Where, Who_affected	What, Where	What, Perpetrator	What, Where
Partial ordering	What < Where Who_affected, What, Where < Why	What < Where	---	What < Where What, Where < Countermeasures, Damages
For the majority of summaries				
In common	---	When, Perpetrator, Why, History	History	When, Prediction
In the 1st paragraph	---	Perpetrator, When	---	Who_affected, When, Damages
Partial ordering	---	---	Who_affected, What, Perpetrator < History	What, Where < Who_affected

For instance, for the summaries of natural disasters, the table shows in the first part (for all summaries) that: all summaries present the *what*, *where*, *who_affected*, *countermeasures* and *damages* aspects in common; *what* and *where* aspects always happen in the 1st paragraph of the summaries; the *what* aspect always appears before (indicated by the symbol <) the *where* aspect and the *what* and *where* aspects always appear before *countermeasures* and *damages* aspects. The second part of the table shows new patterns that arise when we consider the majority of summaries (instead of restricting the analysis to all of the summaries). For instance, still for the natural disasters category, one may see that: *when* and *prediction* aspects appear in the majority of the summaries; *who_affected*, *when* and *damages* aspects happen in the 1st paragraph of the majority of the summaries; and *what* and *where* aspects appear before the *who_affected* aspect.

By the end of the corpus analysis, it was also possible to realize that:

- the *when* aspect does not happen for accidents; the *why* aspect does not happen for natural disasters; the *damages* aspect does not happen for attacks;
- the *what* aspect is usually fragmented in the summaries of legal and political decisions;
- the *why*, *history* and *countermeasures* aspects may appear in paragraphs dedicated to them in the summaries of attacks;
- *history* and *countermeasures* aspects tend to occur in the end of summaries;
- most of the summaries present extra material.

It is important to notice that some of the above considerations are only indicative of summary content and may not be generalized and accepted as usual, since our "world" clusters present few summaries, especially for some categories (such as the accidents category, which has only 2 summaries).

Finally, from the observed data, it may be possible to suggest prototypical structures to compose summaries belonging to "world" section. For instance, the first paragraph ought to contain the aspects *what*, *where* and *who_affected*, in this order. The other aspects probably depend on the categories to which summaries belong.

Final remarks

After the corpus analysis, the annotation team concluded that specific domain knowledge was usually not necessary for the aspect annotation (at least for this "world" section), and only general common sense was used. Domain specific aspects were not necessary as well, since aspects of general use showed to be enough. However, sometimes the annotators had to read the texts that gave origin to the summaries in order to understand some summary passages and to identify the main information.

Our analysis results partially apply to "daily news" section too, since the difference between world and daily news sections appears to be the internationality level of the news under focus. Both sections present accidents, natural disasters and other categories in common.

Acknowledgements

The authors are grateful to FAPESP, CAPES, and CNPq.

References

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)*. Technical Report, Universidade de São Paulo, n° 326, 12p.

CARDOSO, P.C.F.; MAZIERO, E.G.; CASTRO JORGE, M.L.R.; 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: Proceedings of the 3rd RST Brazilian Meeting*, pp. 88-105.

CASTRO JORGE, M.L.R.; DI FELIPPO, A.; NÓBREGA, F.A.A.; SOUZA, J.W.C. (2012). "Analysis of Aspects in a Corpus of Human Multi-document Summaries of Sports News", *in: Anais do XI Encontro de Linguística de Corpus*.

LI, P.; WANG, Y.; GAO, W.; JIANG, J. (2010). "Generating Aspect-oriented Multi-Document Summarization with Event-aspect model", *in: Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1137-1146.

MANI, I. (2001). *Automatic Summarization*. Amsterdam: John Benjamins Publishing Co.

OWCZARZAK, K. and DANG, H.T. (2011). "Who wrote What Where: Analyzing the content of human and automatic summaries", *in: Proceedings of the Workshop on Automatic Summarization for Different Genres, Media, and Languages*, pp. 25-32.

RADEV, D.R. and MCKEOWN, K. (1998). "Generating natural language summaries from multiple on-line sources." *Computational Linguistics*, vol. 24, n° 3, pp. 469-500.

RASSI, A.P.; RINO, L.H.M.; DIAS, M.S. (2012). "Preliminary Aspects Distribution in Political Texts", *in: Anais do XI Encontro de Linguística de Corpus*.

WHITE, M.; KORELSKY, T.; CARDIE, C.; NG, V.; PIERCE, D.; WAGSTAFF, K. (2001). "Multidocument summarization via information extraction", *in: Proceedings of the 1st International Conference on Human Language Technology Research*, pp. 1-7.

ZACARIAS, A.C.I.; AGOSTINI, V.; CARDOSO, P.C.F.; SENO, E.M.R. (2012). "Análise de Aspectos de Sumários Multidocumento e sua Correlação com a Informatividade", *in: Anais do XI Encontro de Linguística de Corpus*.

ZHOU, L.; TICREA, M.; HOVY, E. (2005). "Multi-document Biography Summarization", *in: Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1-8.