

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

## 1. 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 (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.

We report our corpus analysis and results in Section 2. Section 3 presents some final remarks.

## 2. 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 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.

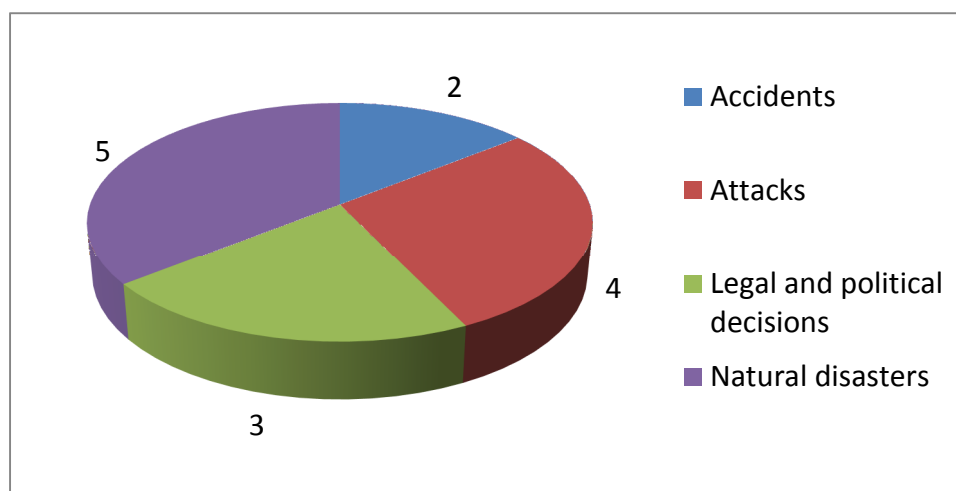


Figure 1. Distribution of categories in the CSTNews “world” section

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). The aspects are shown in capital letters after the text passages they refer to, which are delimited by brackets.

```
[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
```

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.

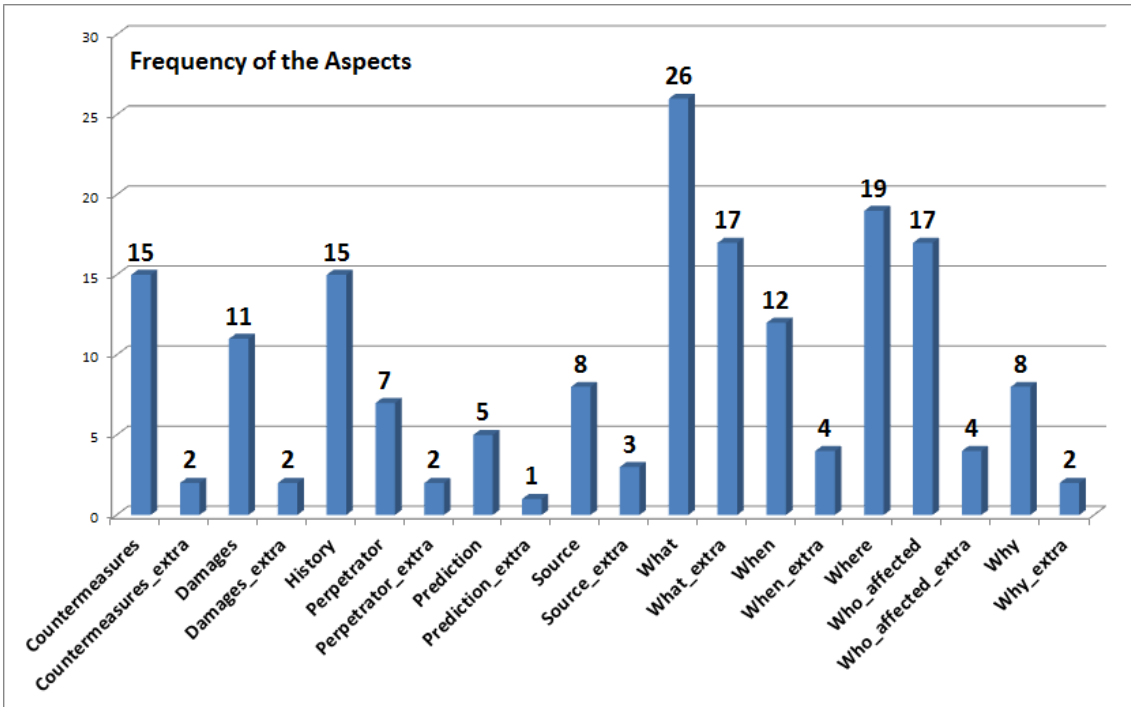


Figure 3. Frequency of aspects in the “world” summaries

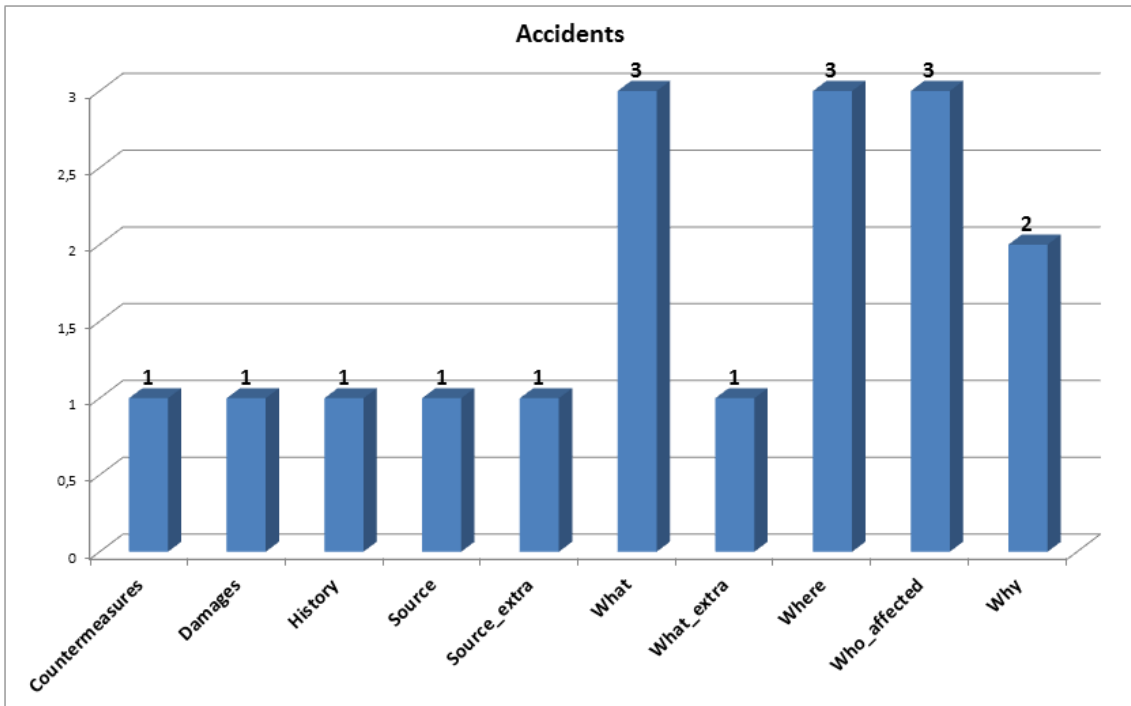


Figure 4. Frequency of aspects in “Accidents” category

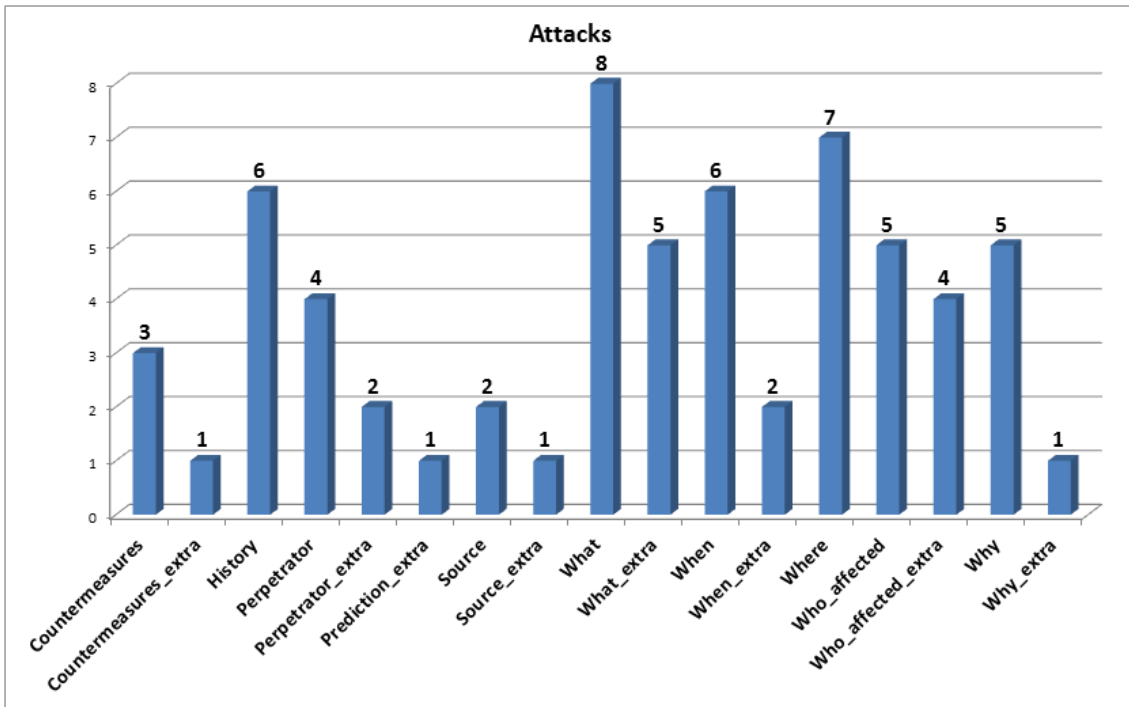


Figure 5. Frequency of aspects in “Attacks” category

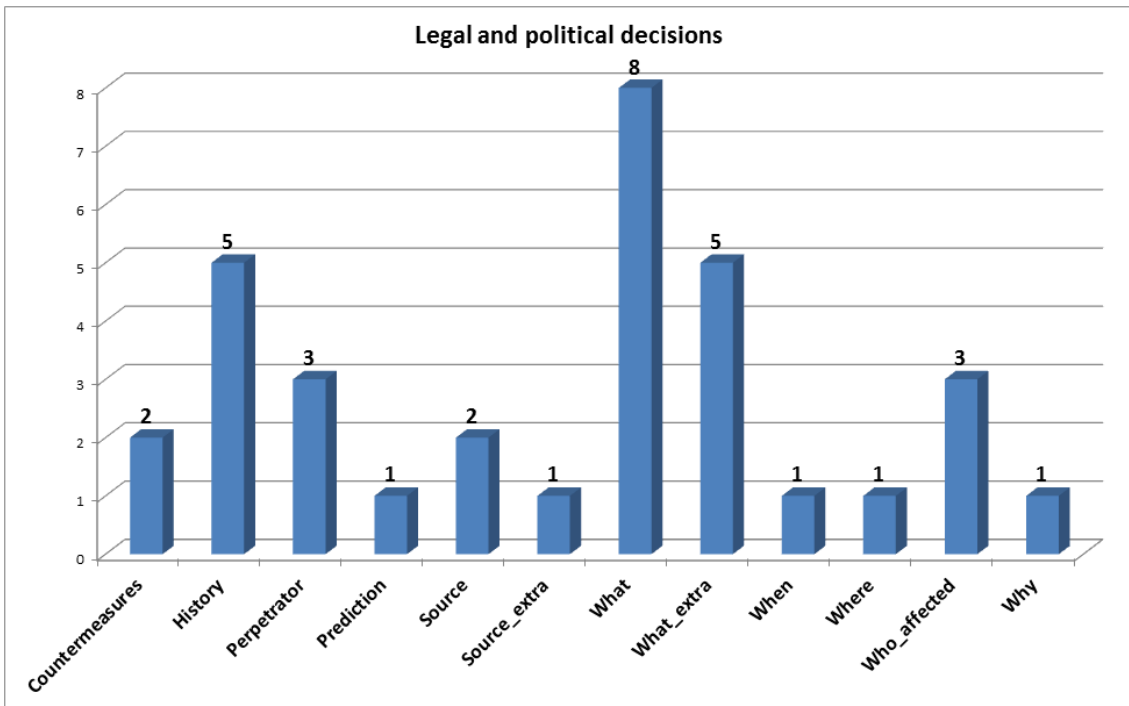


Figure 6. Frequency of aspects in “Legal and political decisions” category

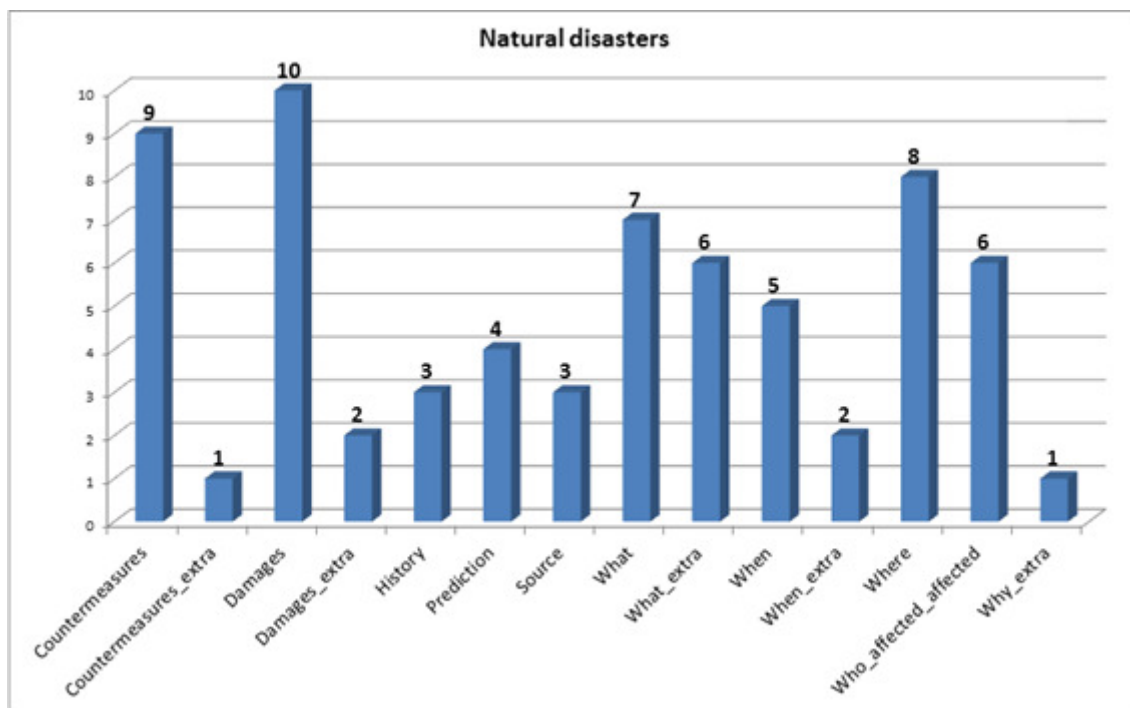


Figure 7. Frequency of aspects in “Natural disasters” category

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
<b>For all summaries</b>				
<i>In common</i>	What, Where, Who_affected, Why	What, Where, Who_affected	What, Perpetrator	What, Where, Who_affected, Countermeasures, Damages
<i>In the 1<sup>st</sup> paragraph</i>	What, Where, Who_affected	What, Where	What, Perpetrator	What, Where
<i>Partial ordering</i>	What < Where Who_affected, What, Where < Why	What < Where	---	What < Where What, Where < Countermeasures, Damages
<b>For the majority of summaries</b>				
<i>In common</i>	---	When, Perpetrator, Why, History	History	When, Prediction
<i>In the 1<sup>st</sup> paragraph</i>	---	Perpetrator, When	---	Who_affected, When, Damage
<i>Partial ordering</i>	---	---	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 1<sup>st</sup> 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 damage aspects happen in the 1<sup>st</sup> paragraph of the majority of the summaries; and what and where aspects appear before the who\_affected aspect.

For better analyzing the ordering of aspects in categories, colored graphics were built. Figure 8 to 11 show the graphics for each of the categories, where the aspects of each summary of the categories are shown in the order they occur in the summaries and similar aspects are linked with colored lines for facilitating the analysis. For instance, one may easily see that the history and countermeasures aspects tend to occur in the end of summaries.

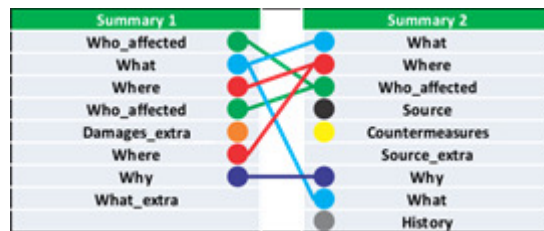


Figure 8. Aspect ordering in “Accidents” category

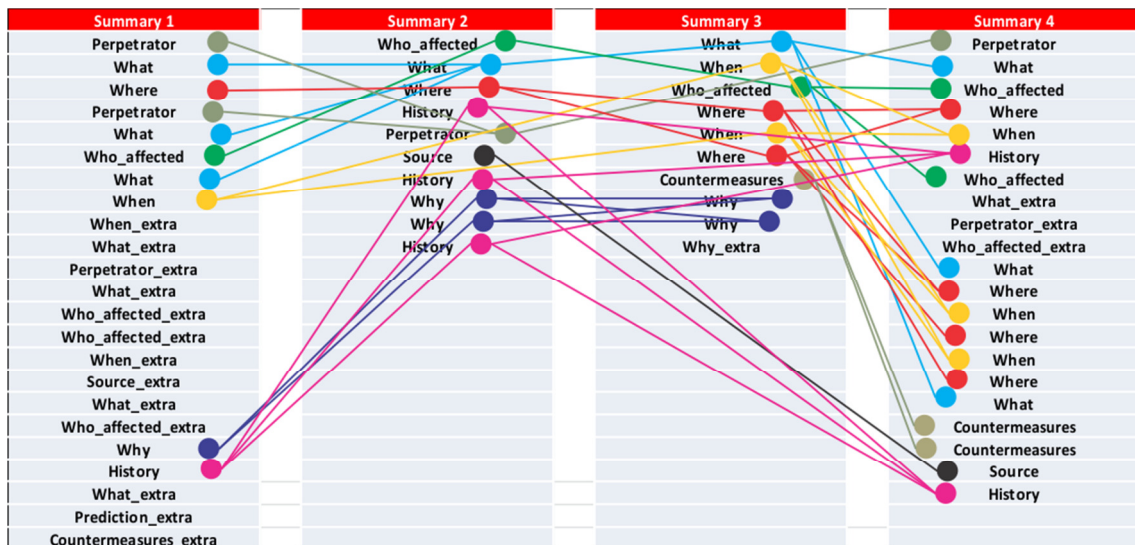


Figure 9. Aspect ordering in “Attacks” category

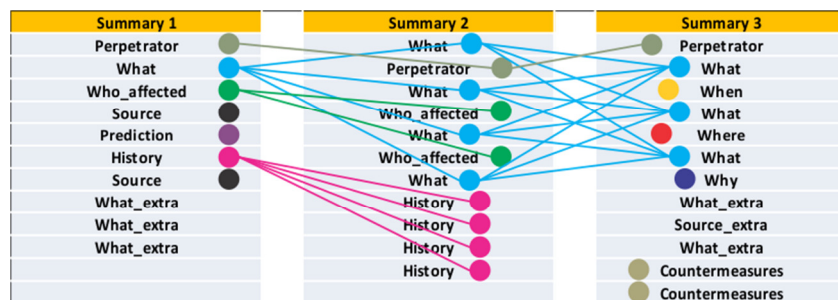


Figure 10. Aspect ordering in “Legal and political decisions” category

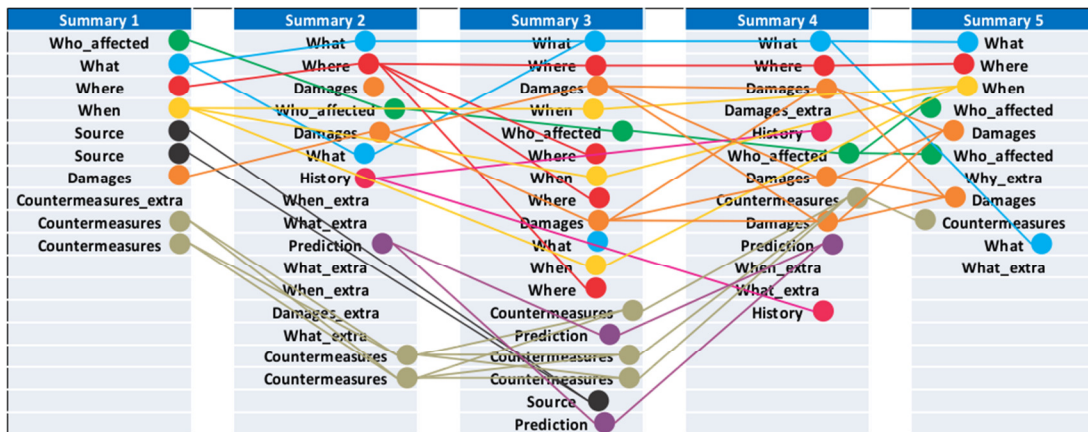


Figure 11. Aspect ordering in “Natural disasters” category

It is also worthy citing some curiosities:

- surprisingly, 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;
- 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 (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.

### 3. Final remarks

After the corpus analysis, the annotation team concluded that specific domain knowledge was 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.

We believe that our analysis results might also apply to “daily news” section, 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.

### References

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.

- Li, P.; Wang, Y.; Gao, W.; Jiang, J. (2010). Generating Aspect-oriented Multi-Document Summarization with Event-aspect model. In the *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1137–1146.
- Mani, I. (2001). *Automatic Summarization*. John Benjamins Publishing Co., Amsterdam.
- Owczarzak, K. and Dang, H.T. (2011). Who wrote What Where: Analyzing the content of human and automatic summaries. In the *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.
- White, M.; Korelsky, T.; Cardie, C.; Ng, V.; Pierce, D.; Wagstaff, K. (2001). Multidocument summarization via information extraction. In the *Proceedings of the 1<sup>st</sup> International Conference on Human Language Technology Research*, pp. 1-7.
- Zhou, L.; Ticea, M.; Hovy, E. (2005). Multi-document Biography Summarization. In the *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1-8.