ANALYSIS OF ASPECTS IN A *CORPUS* OF HUMAN MULTI-DOCUMENT SUMMARIES OF “SPORTS” NEWS

1. Introduction

Multi-Document Summarization (MDS) consists in generating a unique summary from multiple texts on a same event or topic (Mani, 2001). In the last years, there has been much interest in the task of Aspect-Oriented or Guided Summarization from multiple documents, which was introduced by TAC 2010 (Text Analysis Conference), one of many evaluation workshops organized to encourage research in Natural Language Processing.

Different from generic and query-focused summarization, Guided Multi-document Summarization attempts to build a summary by following some pre-defined aspects that characterize a particular topic, according to the user’s needs (Li *et al.*, 2011). We define an aspect as an information unit that commonly appears in texts of a particular category. For example, if we consider a group of texts about “natural disasters”, a summary should cover the following characteristic aspects: what happened, when, why, who was affected, damages and countermeasures (Owczarzak and Dang, 2011). The characterization of aspects requires a deep linguistic (semantic) analysis of human multi-document summaries in order to define general guidelines for their production.

There are a few studies that investigate aspects for multi-document summaries. White et al. (2001) proposed aspect templates for “natural disasters” domain. Afantenos *et al*. (2004) proposed a summarization method using “message templates” and a football ontology, which encapsulated various aspects for the football domain. For instance, they considered aspects such as: entity, performance, time_span, among others. Zhou *et al*. (2005) investigated aspects’ occurrence in biographical summaries. Li *et al*. (2011) explored entity occurrence in Wikipedia summaries, identifying their usual aspects according to the Wikipedia categories. Owczarzak and Dang (2011) studied the impact of aspects in automatic summaries for the “attacks”, “health and safety”, “endangered resources”, and “trials and investigations” categories.

Following the tendencies of many previous works and expecting to contribute to the linguistic characterization of human summaries for future works on MDS, we developed a *corpus*-based analysis of aspects in human multi-document summaries. For this aim, we used the CSTNews *corpus* (Cardoso *et al*., 2011), which is composed by 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*). The texts in CSTNews belong to different categories such as: “world”, “politics”, “sports”, and “daily news”. Each cluster is composed of (i) 2 or 3 texts on the same topic and (ii) automatic and manual versions for both single and multi-document summaries. In particular, we focus on the analysis of aspects for the “sports” category, where we identify content aspects and their organization.

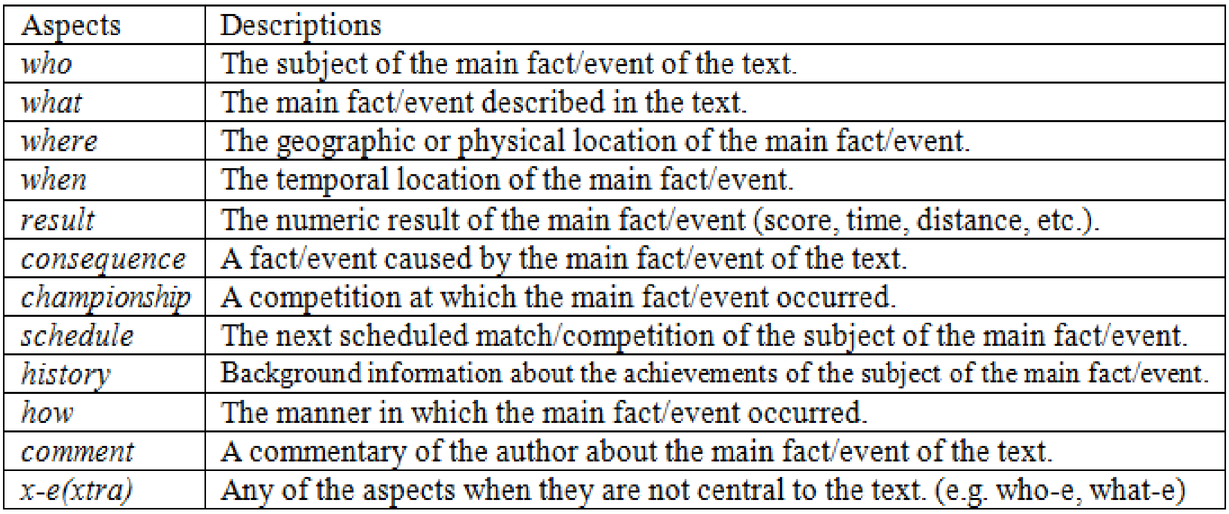
Our *corpus* analysis is described in Section 2. In Section 3, we present the validation of the *corpus* analysis. In Section 4, we present some final remarks.

2. *Corpus* Analysis

Our *corpus* analysis was based on the annotation of aspects for the “sports” category of the CSTNews *corpus*. This category is composed of 10 clusters of journalistic texts, each one containing 2 or 3 texts and the correspondent human multi-document summary. The category covers 2 clusters on swimming competition, 2 on volleyball match, 1 on football match, 1 on pole vault competition, 1 on football and volleyball matches, and 3 clusters on topics that not narrate sports events directly (e.g. Maradona's health). The annotation was performed by 4 annotators together, 2 linguists and 2 computer scientists with certain experience in *corpus* annotation. In order to annotate the occurrence of the aspects, we chose sentence as our unit of analysis.

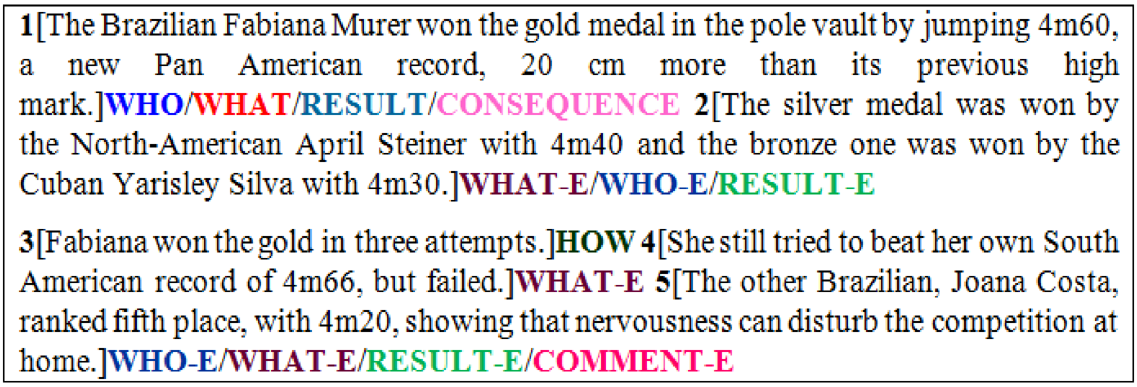
We began our annotation based on the set of generic aspects proposed by the TAC 2010: *who*, *what*, *where*, *when*, and *how* (Owczarzak and Dang, 2011). Under the annotation, we soon identified some more aspects than the generic one. The complete set of the aspects for the "sports" category is listed and defined in the Table 1.

Table 1. List of aspects for "sports" category of the CSTNews *corpus*.



Aspects	Descriptions
who	The subject of the main fact/event of the text.
what	The main fact/event described in the text.
where	The geographic or physical location of the main fact/event.
when	The temporal location of the main fact/event.
result	The numeric result of the main fact/event (score, time, distance, etc.).
consequence	A fact/event caused by the main fact/event of the text.
championship	A competition at which the main fact/event occurred.
schedule	The next scheduled match/competition of the subject of the main fact/event.
history	Background information about the achievements of the subject of the main fact/event.
how	The manner in which the main fact/event occurred.
comment	A commentary of the author about the main fact/event of the text.
x-e(xtra)	Any of the aspects when they are not central to the text. (e.g. who-e, what-e)

All sentences of the human multi-document summaries were annotated according to the aspects listed in Table 1. Each sentence could be associated to one or more aspects, if required. As illustration, Figure 1 shows an annotated summary. The aspects are shown in capital letters in the end of sentence, which is delimited by brackets and numbered in sequence. The sequence of tags follows the aspects sequence occurrence in the sentence.



1[The Brazilian Fabiana Murer won the gold medal in the pole vault by jumping 4m60, a new Pan American record, 20 cm more than its previous high mark.]**WHO/WHAT/RESULT/CONSEQUENCE** **2**[The silver medal was won by the North-American April Steiner with 4m40 and the bronze one was won by the Cuban Yarisley Silva with 4m30.]**WHAT-E/WHO-E/RESULT-E**

3[Fabiana won the gold in three attempts.]**HOW** **4**[She still tried to beat her own South American record of 4m66, but failed.]**WHAT-E** **5**[The other Brazilian, Joana Costa, ranked fifth place, with 4m20, showing that nervousness can disturb the competition at home.]**WHO-E/WHAT-E/RESULT-E/COMMENT-E**

Figure 1. Example of annotated summary.

In Figure 1, the main information is that an athlete won the gold medal with a new record (*what* aspect). The *what_e(xtra)* in the sentence 5, for instance, indicates additional information (the award of the silver medal). Figure 2 shows the overall frequency of the aspects in the "sports" clusters and their frequency in different summaries.

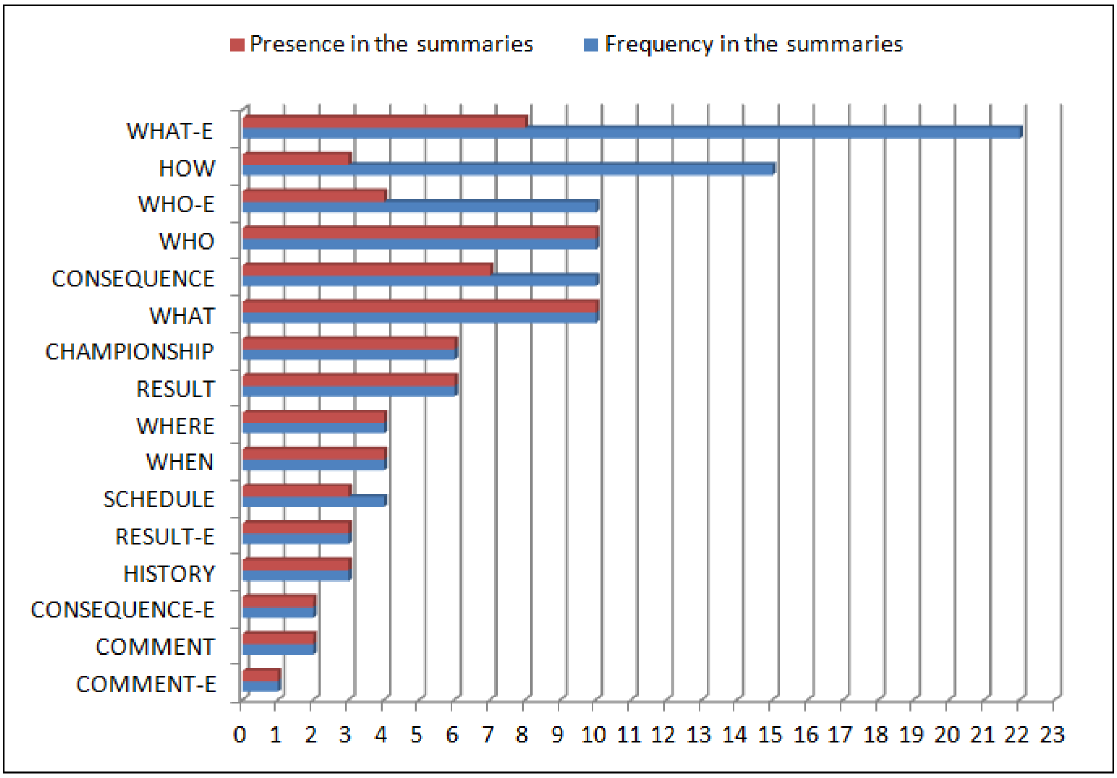


Figure 2. Frequency of aspects in the *corpus*.

It can be observed from Figure 2 that *what-e* and *how* were the most frequent aspects, occurring 22 and 15 times, respectively. This statistic reveals a common pattern in sports news texts, where information extra, different from the main event, is almost always included, and details on how the main event took place are described (e.g., the player who made a goal). Despite its overall frequency, the *how* aspect just occurred in 3 summaries, 2 of them describing football matches. Following the scale, we have *who-e*, *who*, *consequence*, and *what* occurring 10 times, and *championship* and *result* occurring 6 times. Unlike the *how* aspect, *who*, *consequence*, *what*, *championship*, and *result* are very frequent in our *corpus* and they are present in most summaries.

Another observation is that the most frequent and common aspects (i.e. *who*, *consequence*, *what*, *championship*, and *result*) tend to appear more in the first paragraph of the summaries (cf. Figure 1). Figure 3 illustrates the frequency of aspects occurrence in contrast with the frequency of aspects occurrence in the first paragraph.

We also noticed a pattern in the order of aspects occurrence. In other words, we conclude that some aspects are more usual than others and that there are partial orderings among some of them. Specifically, the sequence *who/what* occurred in all summaries while *who/what/consequence* appeared in 7 texts. Others sequences can be seen in Figure 4.

The aspects, however, do not always occur in a direct order, but it was possible to identify partial orderings. For instance, the *who* aspect always appears before (indicated by the symbol <) the *what* aspect, and *who* and *what* always appear before *championship*. The partial orderings are described in Table 2.

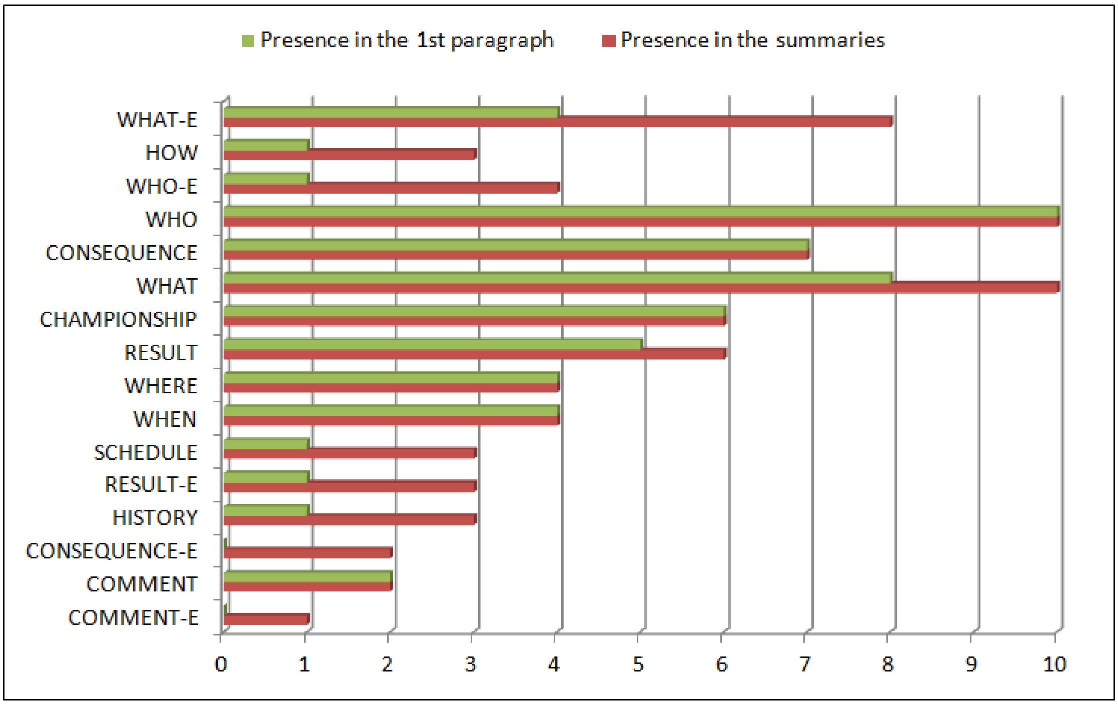


Figure 3. Frequency of aspects in the *corpus* vs. Frequency of aspects in the 1st paragraph.

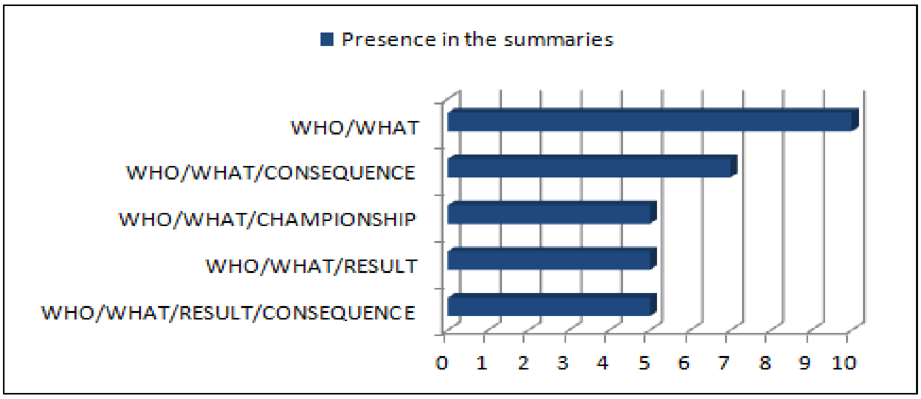


Figure 4: Frequency of aspects sequence occurrence in the *corpus*.

Table 2. Aspect distribution in the summaries.

For all summaries	
In common	*who, what*
In the 1st paragraph	*who, what*
Ordering	*who, what*
For the majority of summaries	
In common	*who, what, result, consequence, championship, what-e*
In the 1st paragraph	*who, what, result, consequence, championship*
Partial ordering	*who < what* *who, what < championship* *result < consequence* *who, what < result, consequence*

In order to visualize better the patterns, we show Table 3. Each column shows the order in which aspects appear in the paragraphs of each summary. Aspects are associated to a particular color in the table, and *x-e(xtra)* aspects are presented in white letters.

Table 3. Resume of aspect occurrence in the *corpus* of sports summaries.

		Summaries									
		Volleyball	Swimming	Swimming	Pole Vault	Volleyball/ Football	Football	Volleyball	Olympic Torch	Fan's reaction	Maradona's Health
Paragraphs	1	who	who	when	who	comment	who	comment	who	when	who
		what	what	who	what	who	what	who	what	champ	what
		result	when	what	result	what	where	what		where	when
		where	champ	result	conseq	champ	how	result		who	conseq
		conseq	what-e	conseq	what-e			where		what	what-e
		champ	result	conseq	who-e			conseq		what-e	
		schedule	conseq	champ	result-e			champ			
								history			
	2	conseq	who-e	who-e	how	conseq	how	what-e	what-e	who-e	what-e
		schedule	what-e	what-e	what-e		how		schedule	what-e	
			conseq-e	who-e	who-e					result-e	
				what-e	what-e					who-e	
				conseq-e	result-e					what-e	
					comment-e					what-e	
	3	history		who-e		conseq	how			who-e	what-e
				what-e		result				what-e	
		what-e		who-e		how				what-e	
				what-e		history				what-e	
	4					how	how			schedule	
						how	how				
							how				
	5						how				
							how				
	6						how				
	7						how				

It is also worthy citing to mention some curiosities:

- The sports category of the CSTNews is actually composed of 7 summaries on sporting events; 3 of the 10 summaries do not describe effectively sports events;
- The *result* aspect does not appear in these 3 summaries as well as the *who/what/result* ordering;
- The *result* and *consequence* aspects did not occur in only 1 summary of the 7 on sporting event;
- The *result* occurs after *consequence* aspect in 1 summary of the 6 in which they appears;
- The *how* aspect is very frequent in texts on football matches.

3. Validation

After the *corpus* annotation, we performed a validation process of our list of aspects in other texts, different from the texts of CSTNews. For this, we built a small test *corpus* composed of 5 clusters of news on the following sports: Football, Volleyball, Tennis, Basketball and Swimming. For each cluster, composed of 2 texts, it was produced a summary by graduate and undergraduate students of different courses. These summaries were annotated by the same 4 annotators according to the list of aspects in Table 1. After that, we computed the frequency of occurrence of the aspects in the new *corpus* and the presence of aspects in each summary (Figure 5). It was also computed the frequency of aspects within the first paragraph of each summary (Figure 6).

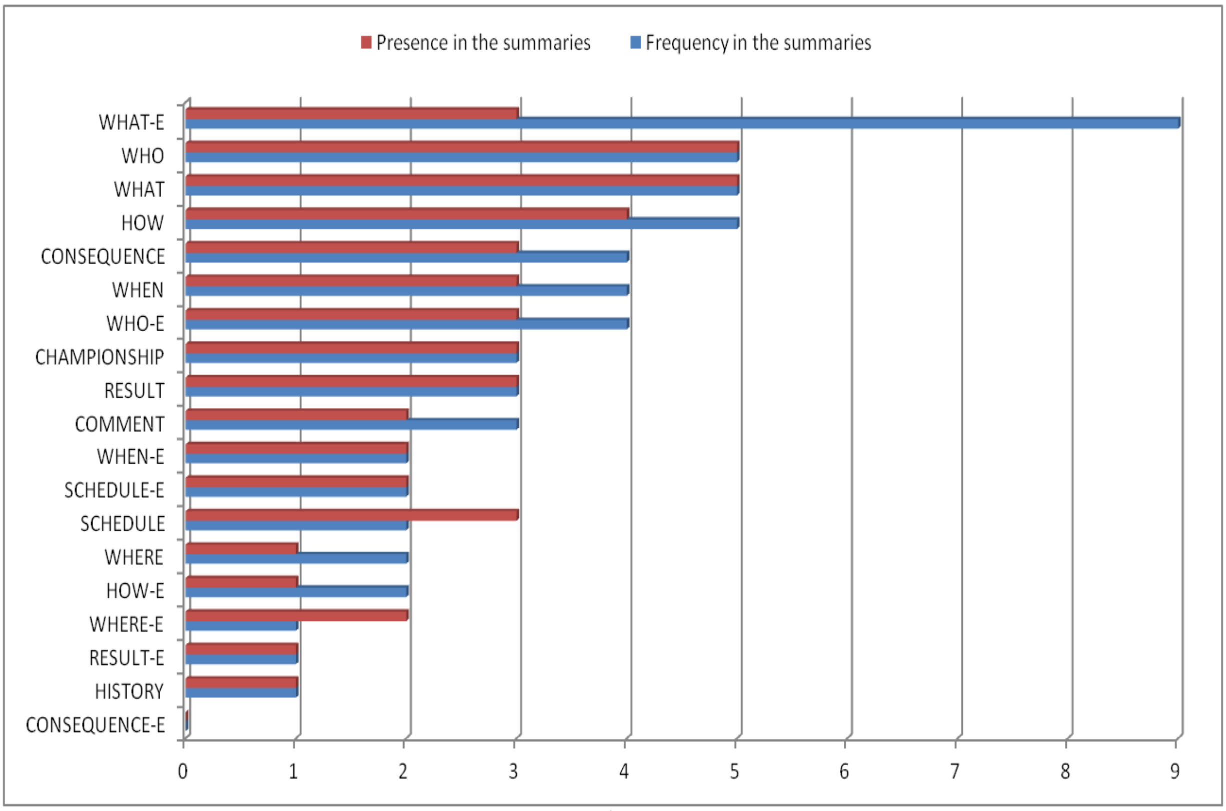


Figure 5. Frequency of aspects in the test *corpus*.

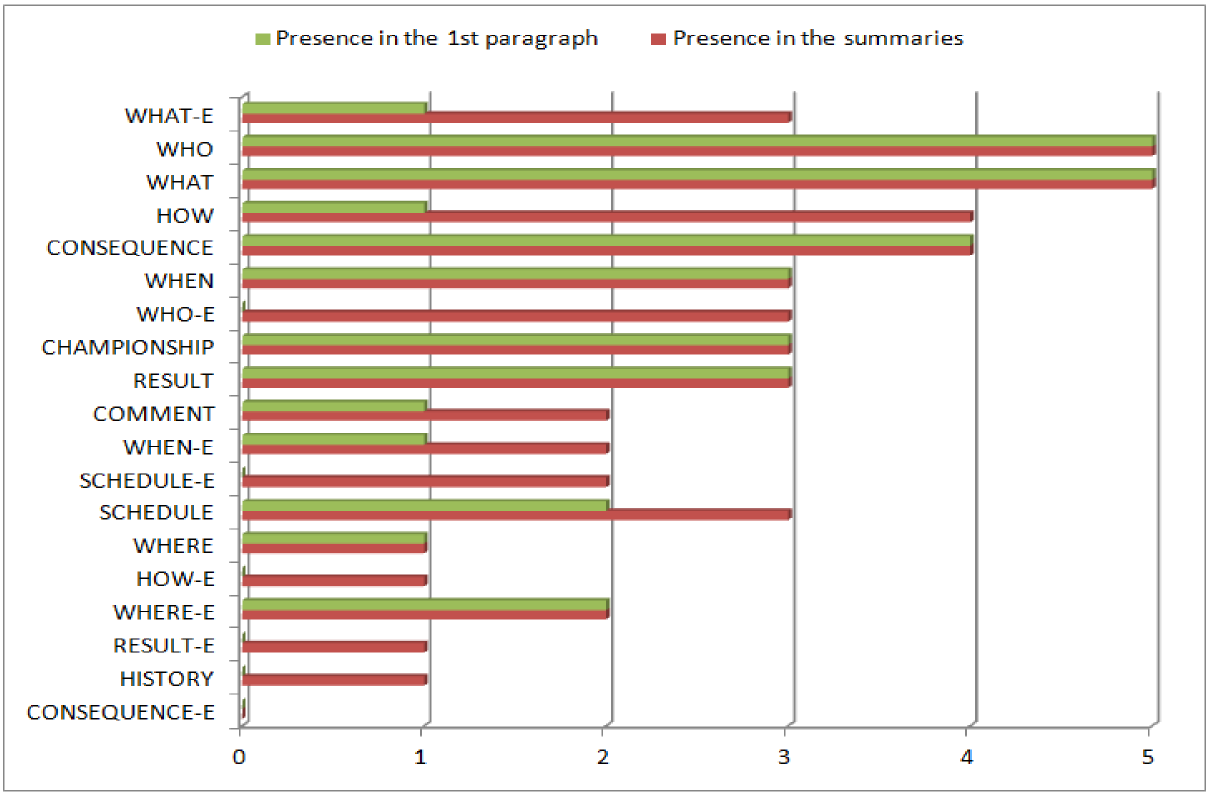


Figure 6. Frequency of aspects in the test *corpus* vs. Frequency of aspects in the 1st paragraph.

We also computed the frequency of occurrence of sequences of aspects in the new summaries. The resume of these results are shown in Figure 7.

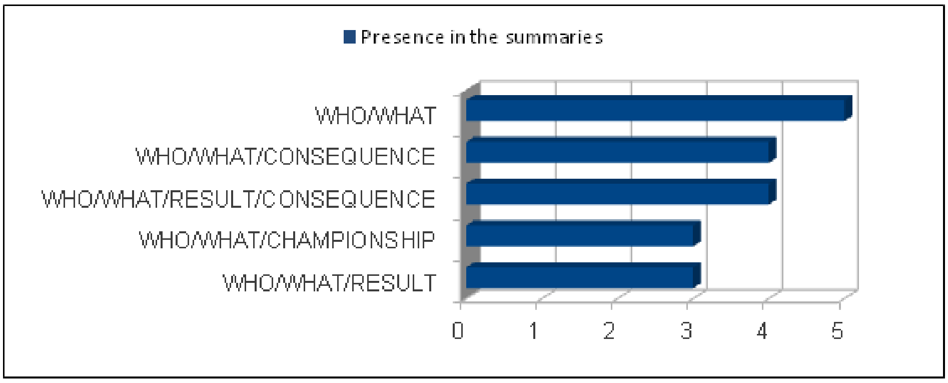


Figure 7. Frequency of aspects sequence occurrence in the test *corpus*.

Based on Figures 5, 6, and 7, we see that the set of aspects, and their distribution and ordering identified in the CSTNews summaries were maintained in the test *corpus*. However, it is important to notice that the previous considerations are only indicative of summary content, since our "sport" clusters from CSTNews present few summaries.

4. Final Remarks

After the *corpus* analysis, we concluded that specific domain knowledge was necessary for the aspect annotation (at least for the *schedule* aspect). Besides, it may be possible to suggest prototypical structures to compose summaries belonging to "sports" section. For instance, the first paragraph ought to contain *who*, *what*, *result*, *championship* and *consequence* aspects, in this order.

References

Afantenos, S. D., Doura, I., Kapellou, E., and Karkaletsis, V. Exploiting cross-document relations for Multi-document Evolving Summarization. (2004). In the *Proceedings of 3rd Helenic Conference on AI (SETN 2004). LNAI 3025*, pp. 410-419.

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. (2011). 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 1st International Conference on Human Language Technology Research*, pp. 1-7.

Zhou, L.; Ticrea, M.; Hovy, E. (2005). Multi-document Biography Summarization. In the *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1-8.