Analysis of Aspects in a *Corpus* of Human Multi-document Summaries of "Sports" News

[1,2] Maria Lucía Castro Jorge
[1,3] Ariani Di Felippo
[1,2] Fernando Antônio Asevedo Nobrega
[1,3] Jackson Wilke da Cruz Souza

1 Núcleo Interinstitucional de Linguística Computacional (NILC)
2 Instituto de Ciências Matemáticas e de Computação (ICMC), Universidade de São Paulo (USP)
3 Departamento de Letras (DL), Universidade Federal de São Carlos (UFSCar)

Schedule

- Context and Motivation
- Goals
- Corpus Analysis
- Validation
- Final Remarks

Introduction: Context and Motivation

- **Multi-document Sumarization (MDS)** has become a very important research area
  - Large collections of data available
  - Many textual data related to a **same topic**
  - Many **phenomena** present (redundancies, complementar information, contradictions, etc.)
  - Sumaries from these groups of texts have become a usefull resource

Introduction: Context and Motivation

- Many approaches for MDS
  - Sentence positition, word frequency, bag of words, cross-document approaches (e.g. Cross-document Structure Theory), among others
  - Recently **Aspect Oriented** **or** **Guided Sumarization**
    - TAC 2010 (Text Analysis Conference)
    - Attempt to build summaries by following pre-defined **aspects**

Introduction: What are “Aspects”?

- Some information units commonly appear in texts related to a same topic, for example:
  - Texts about “natural disasters” include ***what happened***, ***when***, ***why***, ***who was affected***, ***damages*** and ***countermeasures*** (Owczarzak and Dang, 2011)
- These information units are called ***aspects***
- The aspects are important information to understand the specific content of a document

Introduction: Goals of this work

- **General**
  - Contribution to the **linguistic characterization** of human or manual summaries

- **Specific**
  - Analysis of **aspects** in human multi-document summaries
  - In particular, for this analysis we consider summaries from the **"sports" category** of the CSTNews corpus (Cardoso et al., 2011)

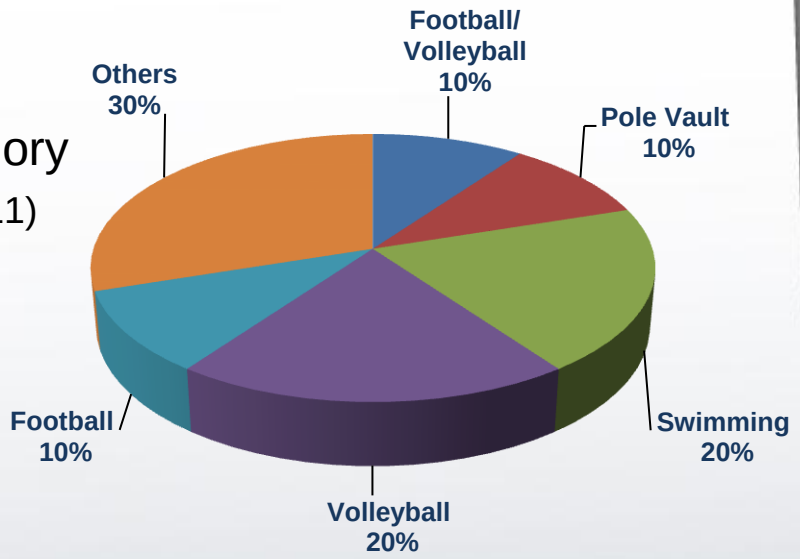
Methodology

- Corpus Analysis
  - Definition of **Aspects for "Sports"** Category
    - Based on the aspects proposed in TAC 2010
  - Statistics of Aspects' ocurrence
- Validation of Aspects
  - **Anotation of 5 new summaries** according to the defined aspects
  - Statistics for the new anotation
    - **Do this validate our set of Aspects?**

Corpus analysis

Corpus

- Manual summaries of the “sports” category of the CSTNews *corpus* (Cardoso *et al.*, 2011)
  - **10** clusters



Annotation team

- 2 linguists and 2 computer scientists

Initial guidelines

- Sentence as unit of analysis
- Generic aspects (TAC´2010): *who, what, where, when, how*
- Annotation was done by the 4 annotators together

Corpus analysis

- **Aspects** for “sports” category of the CSTNews

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)

Corpus analysis

- **Aspects** for “sports” category of the CSTNews

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)

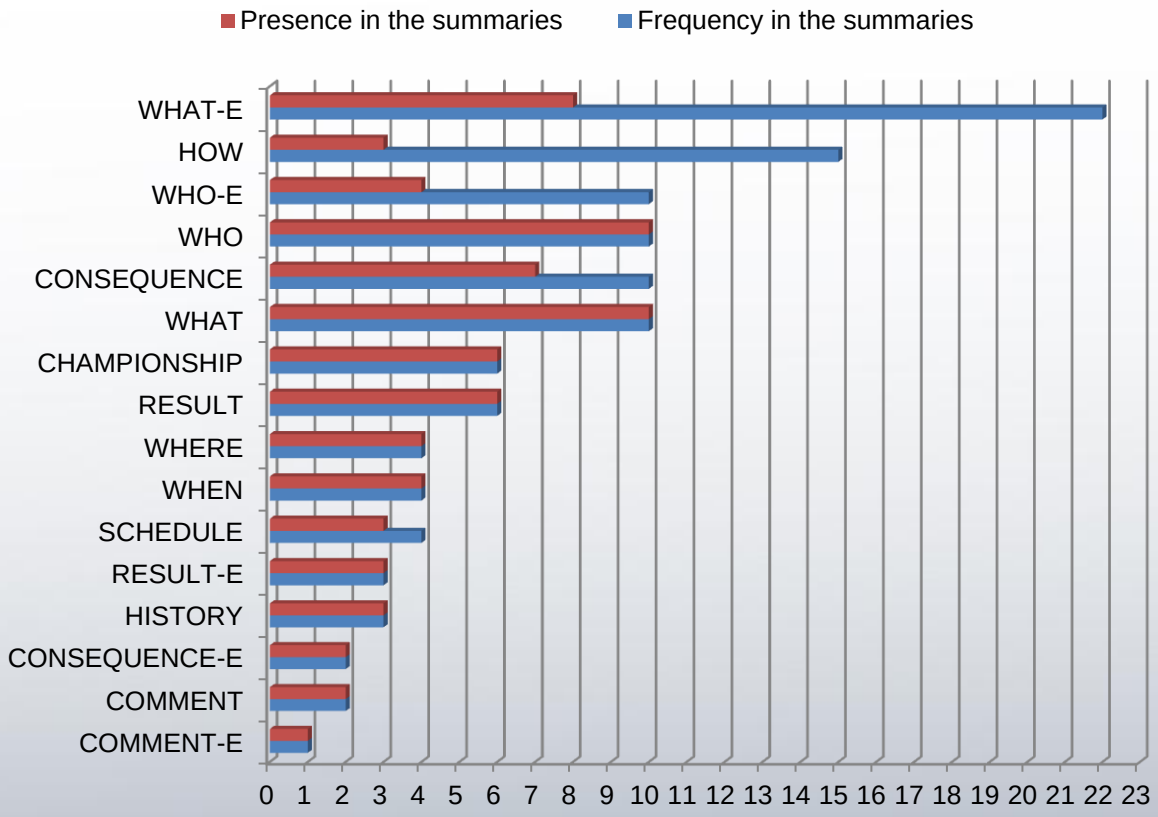
Corpus analysis

- Example of **annotated summary**

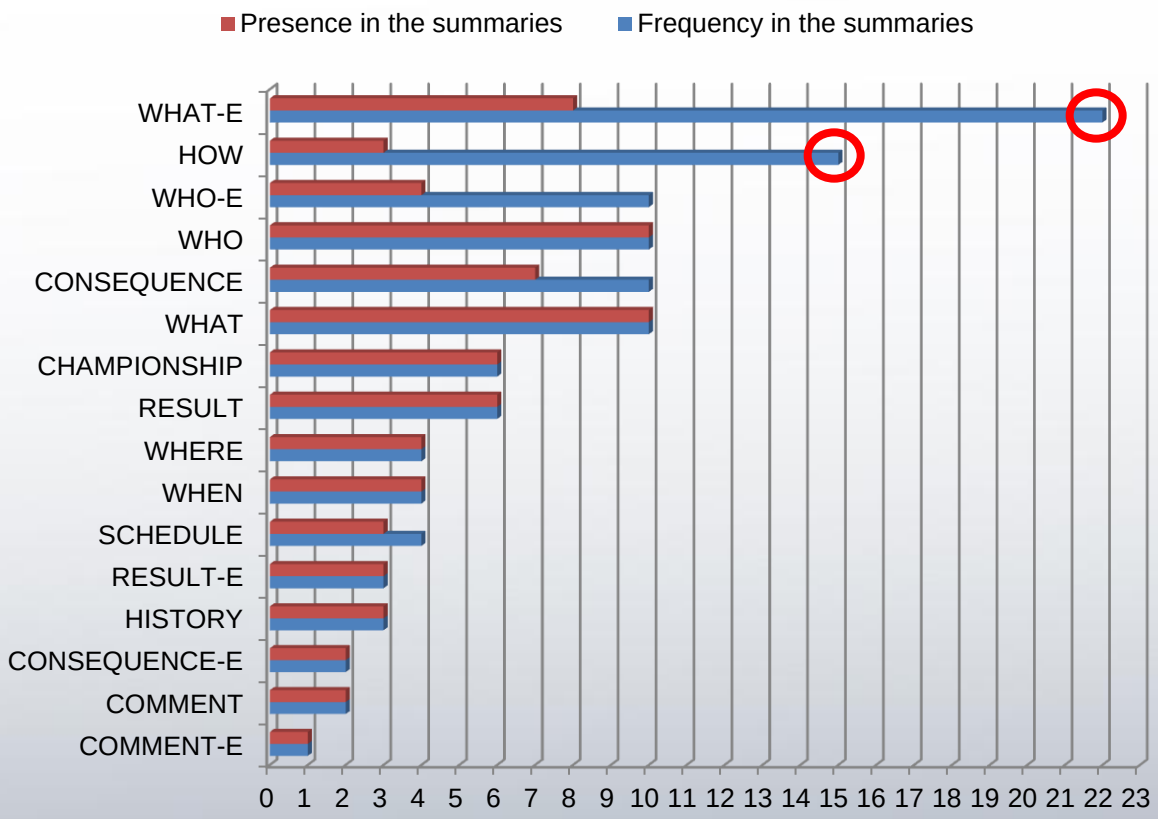
1[A brasileira Fabiana Murer conquistou a medalha de ouro no salto com vara ao saltar 4m60, um novo recorde pan-americano, 20 cm a mais que sua antiga marca.**]WHO/WHAT/RESULT/CONSEQUENCE 2[**A medalha de prata ficou com a americana April Steiner com 4m40 e a de bronze com a cubana Yarisley Silva com 4m30.**]WHAT-E/WHO-E/RESULT-E**

3[Fabiana conseguiu o ouro em três tentativas.**]HOW 4[**Tentou ainda bater o próprio recorde sul-americano de 4m66, mas não conseguiu.**]WHAT-E** 5[A outra brasileira, Joana Costa, ficou na quinta posição, com 4m20, mostrando que o nervosismo pode atrapalhar as competições em casa.**]WHO-E/WHAT-E/RESULT-E/COMMENT-E**

Corpus analysis results

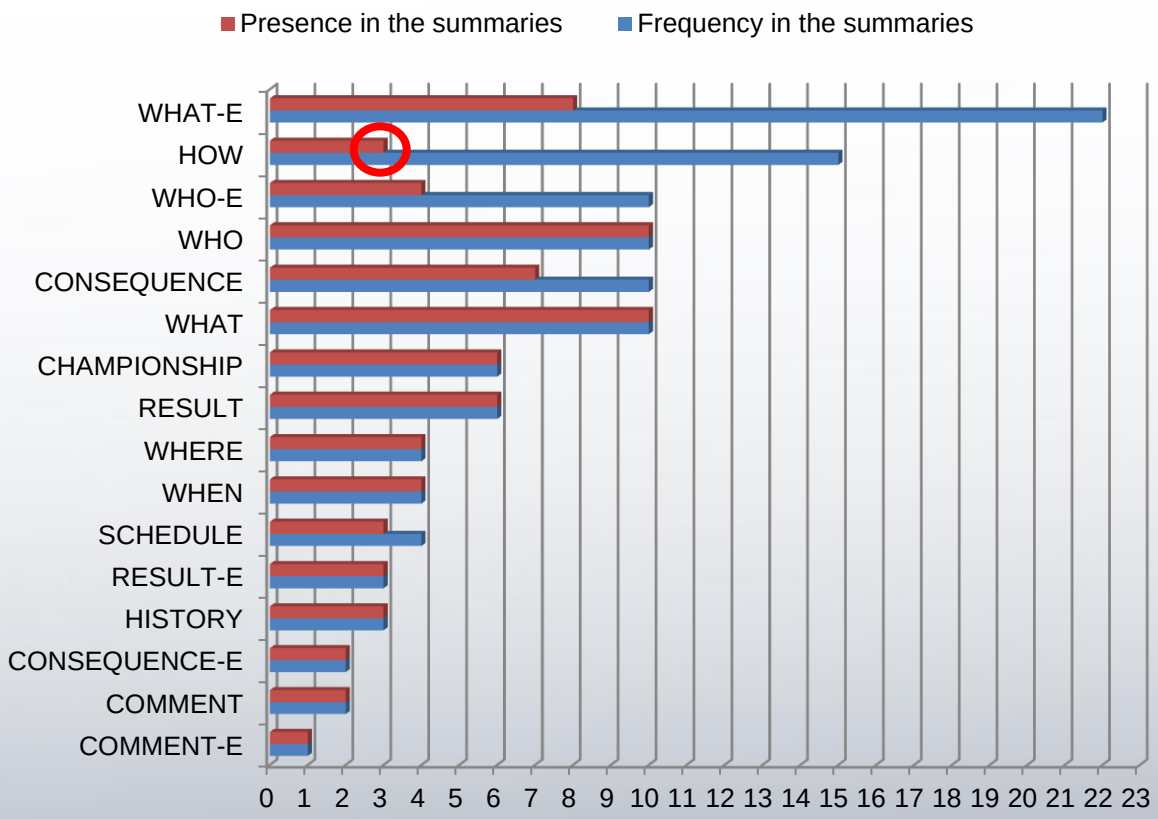


Corpus analysis results



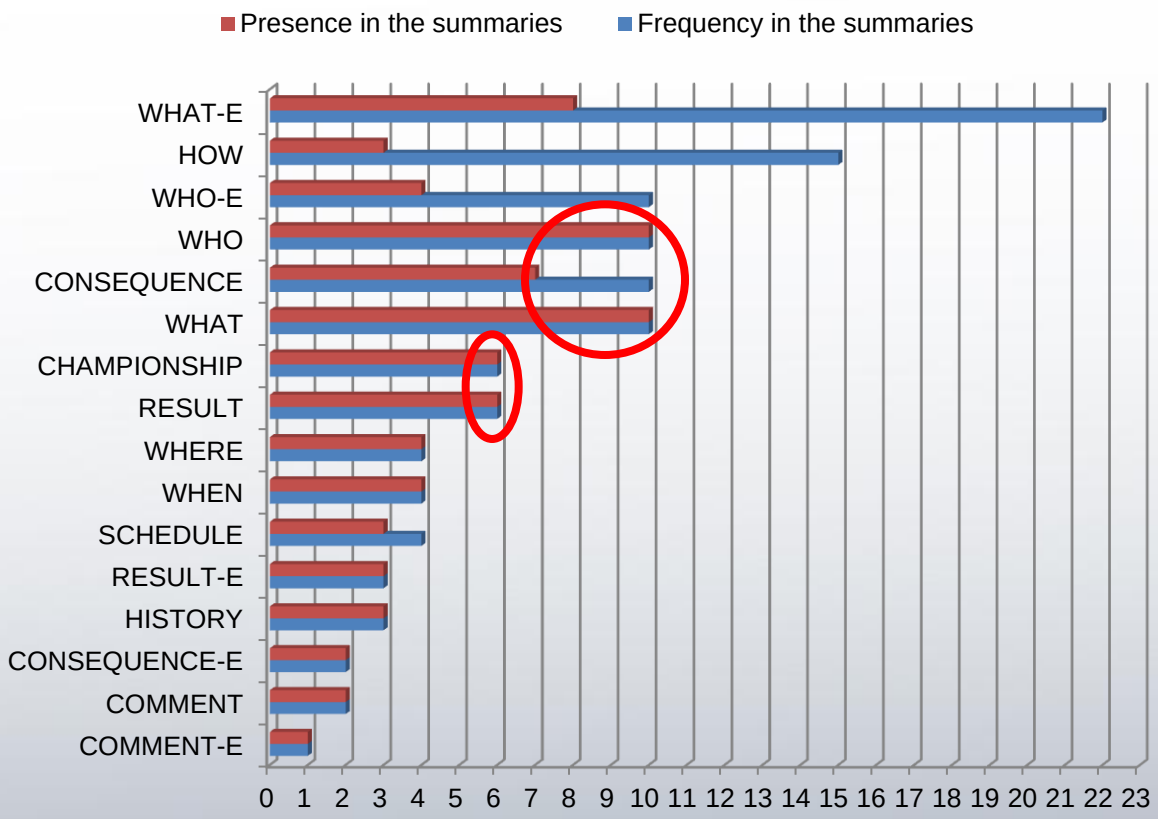
- ***What-e*** and ***how*** are the most frequent aspects
  - Information extra
  - Details on how the main event took place

Corpus analysis results



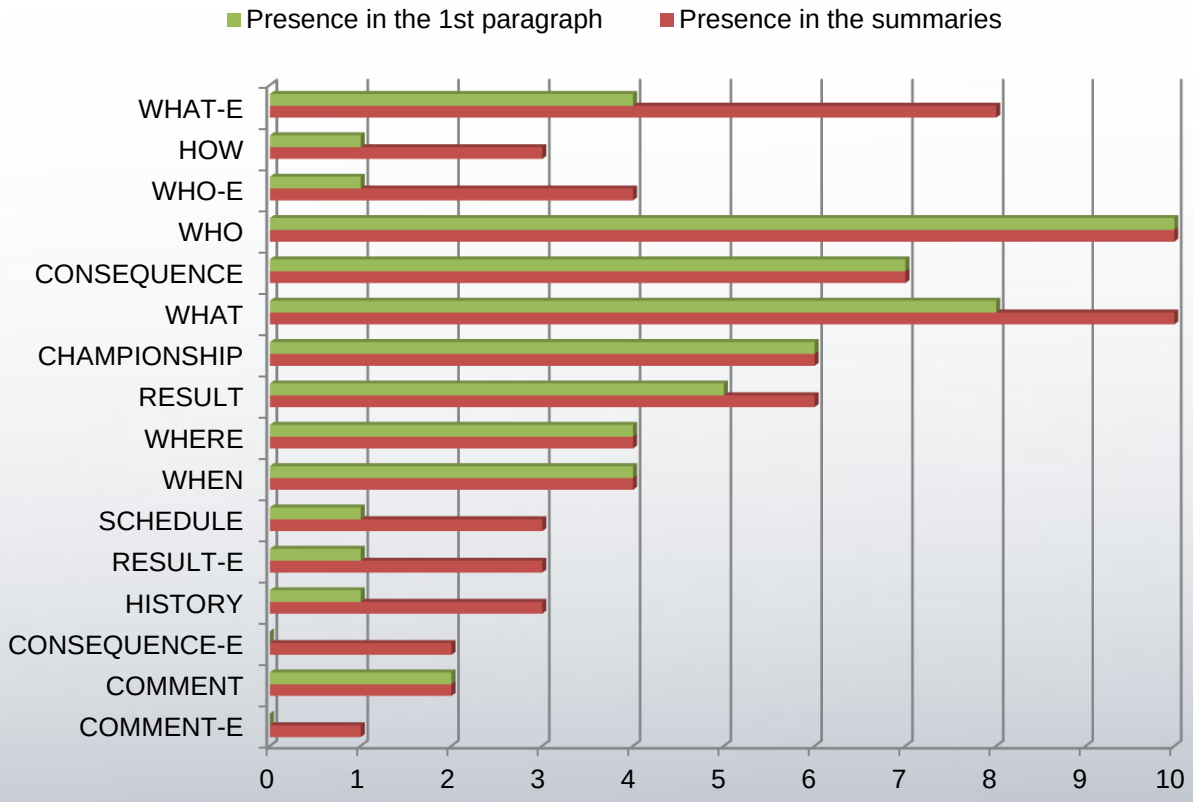
- ***What-e*** and ***how*** are the most frequent aspects
  - Information extra
  - Details on how the main event took place
- ***How*** occurred in 3 summaries (2 on football)

Corpus analysis results

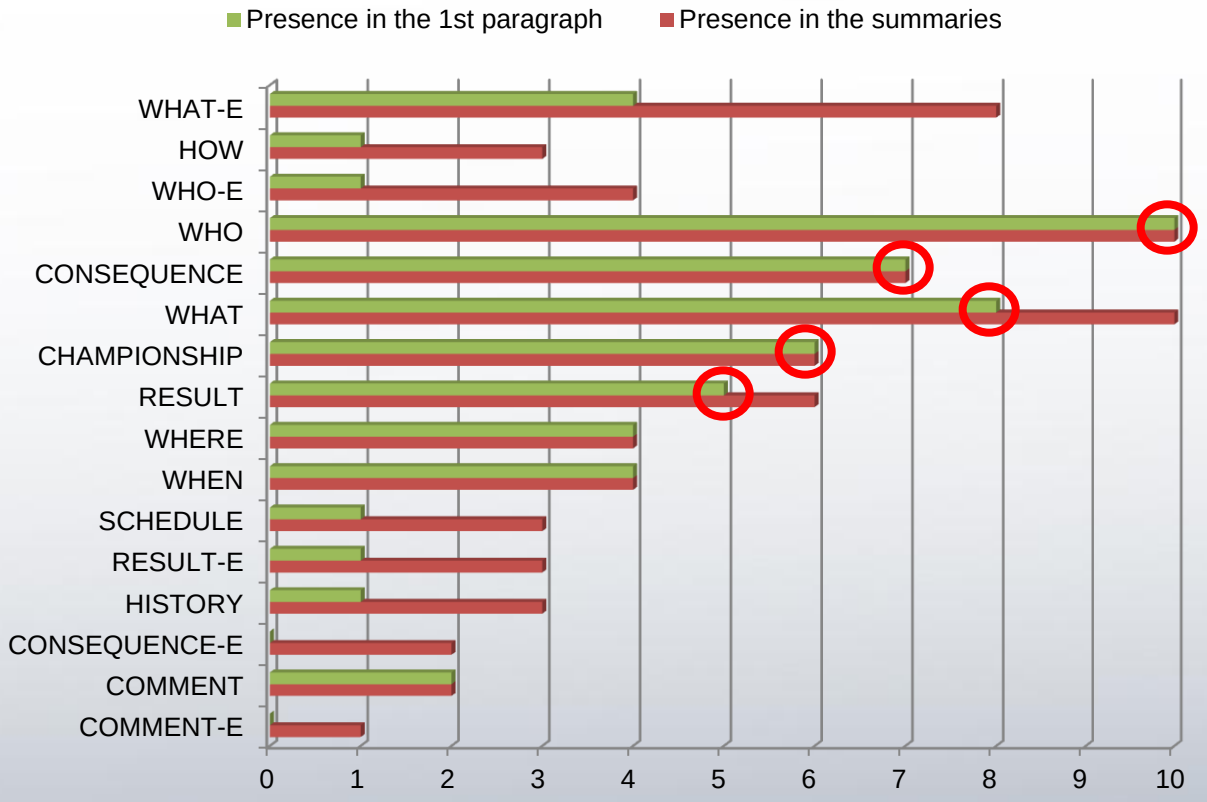


- ***What-e*** and ***how*** are the most frequent aspects
  - Information extra
  - Details on how the main event took place
- ***How*** occurred in 3 summaries (2 on football)
- ***Who***, ***consequence***, ***what***, ***championship***, and ***result*** are very frequent and they are present in most summaries

Corpus analysis results

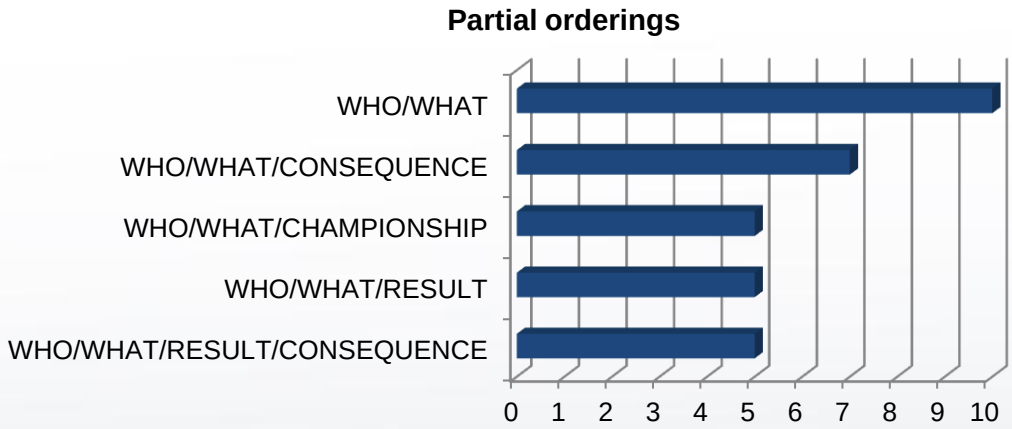


Corpus analysis results



- ***Who***, ***consequence***, ***what***, ***championship***, and ***result***
  - Most frequent in 1st paragraph

Corpus analysis results



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*

		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				

		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				

who, **what** < **result**, **consequence**

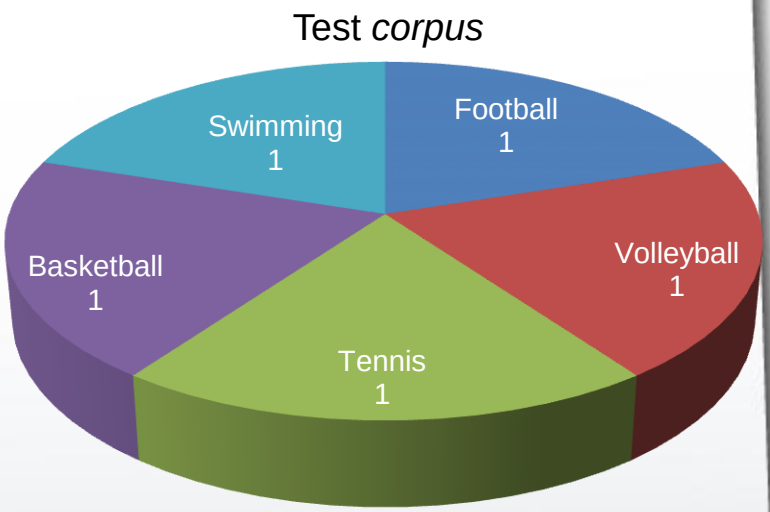
who, **what** < **championship**

Corpus analysis results

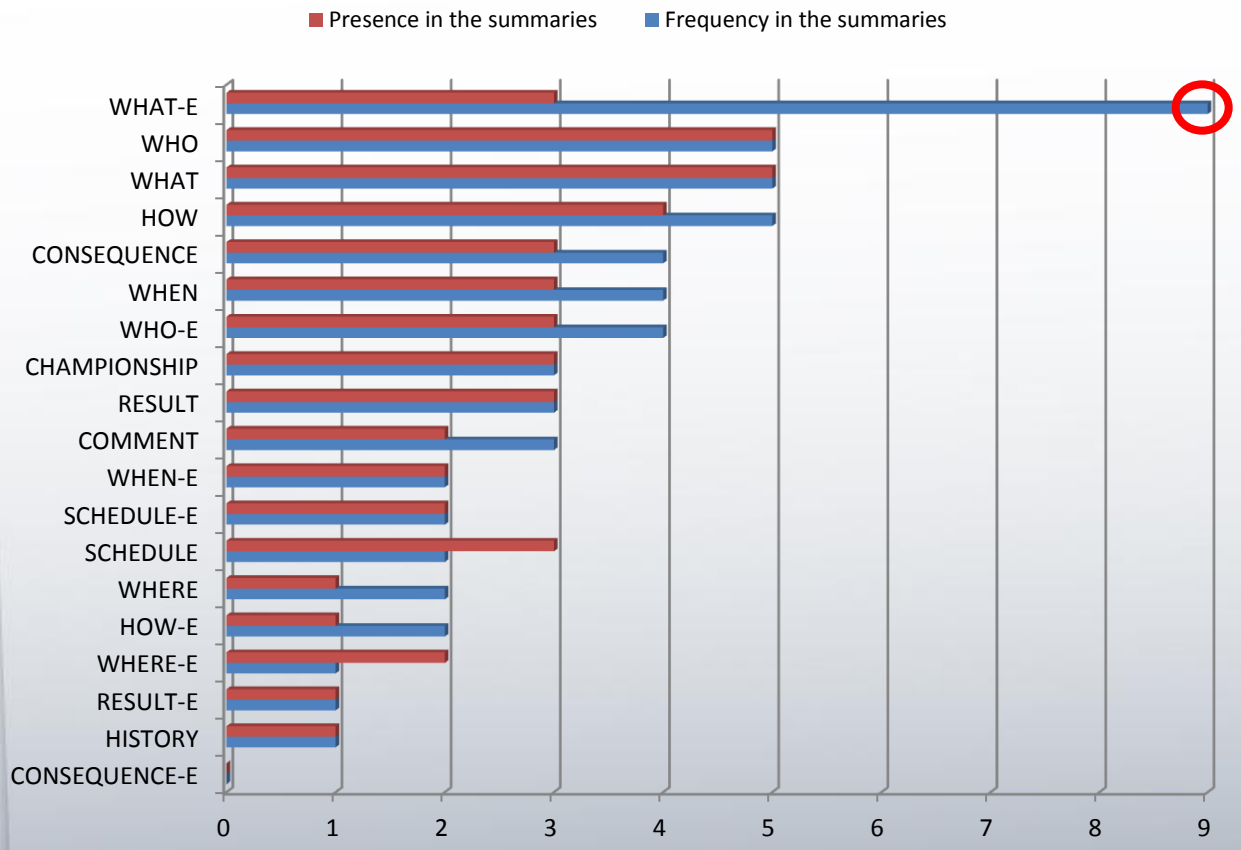
- Some curiosities:
  - The sports category of the CSTNews is actually composed of 7 summaries on sporting events; 3 of the 10 summaries do not describe effectively sports events
  - ***Result*** does not appear in these 3 summaries as well as the *who/what/result* ordering
  - ***Result*** and ***Consequence*** did not occur in only 1 summary of the 7
  - ***Result*** occurs after ***Consequence*** in 1 summary out of the 6 in which they appear
  - **How** is very frequent in texts on football matches

Validation

- Construction of a “test corpus”
  - 5 clusters
    - 2 texts each cluster
    - Manual summaries were produced by graduate and undergraduate students of different courses
    - The summaries were annotated 4 annotators

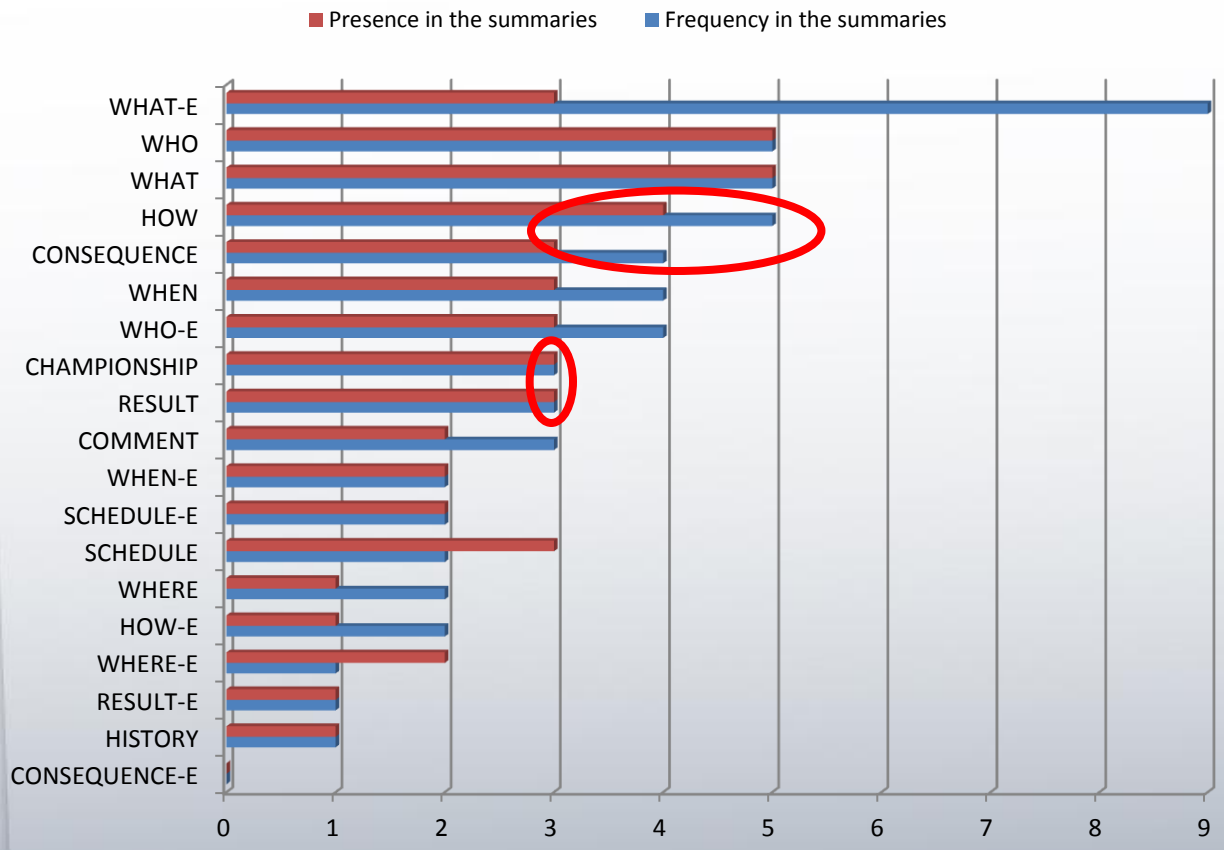


Validation results



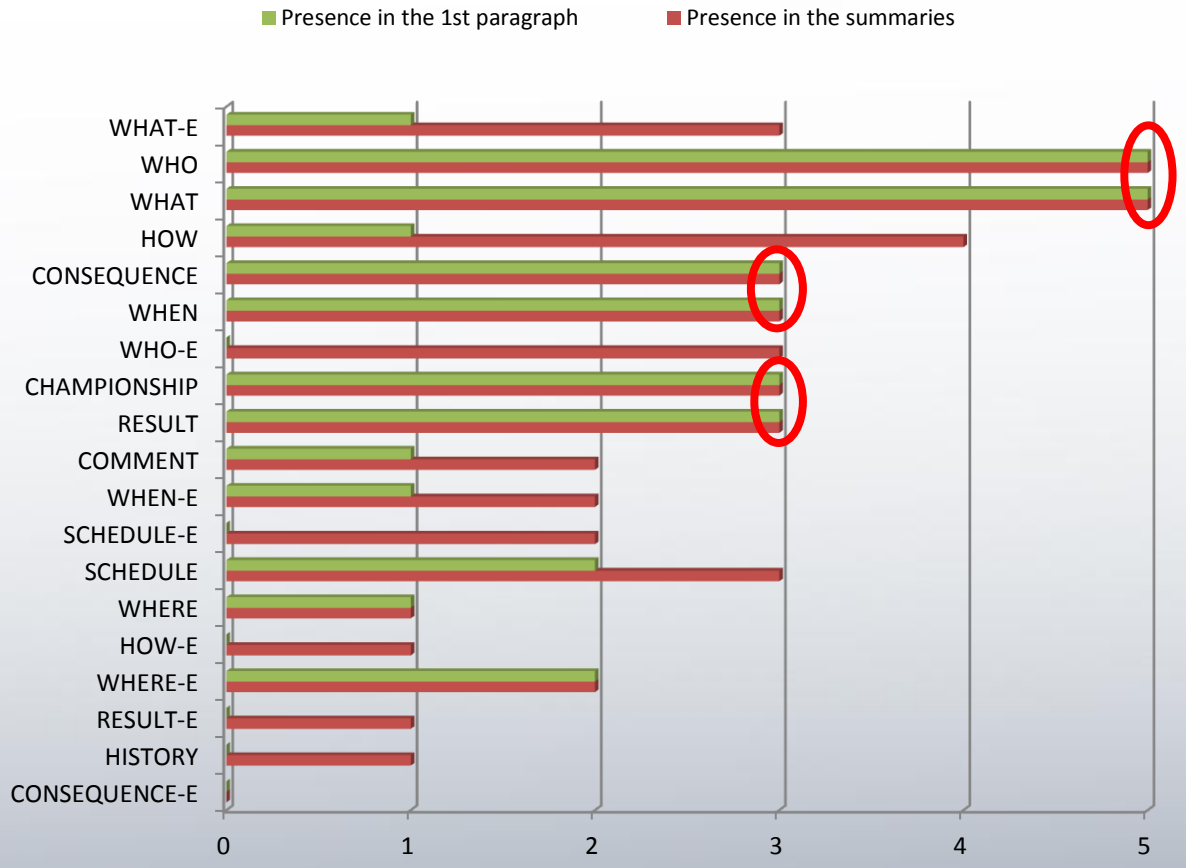
- ***What-e*** is the most frequent aspect

Validation results



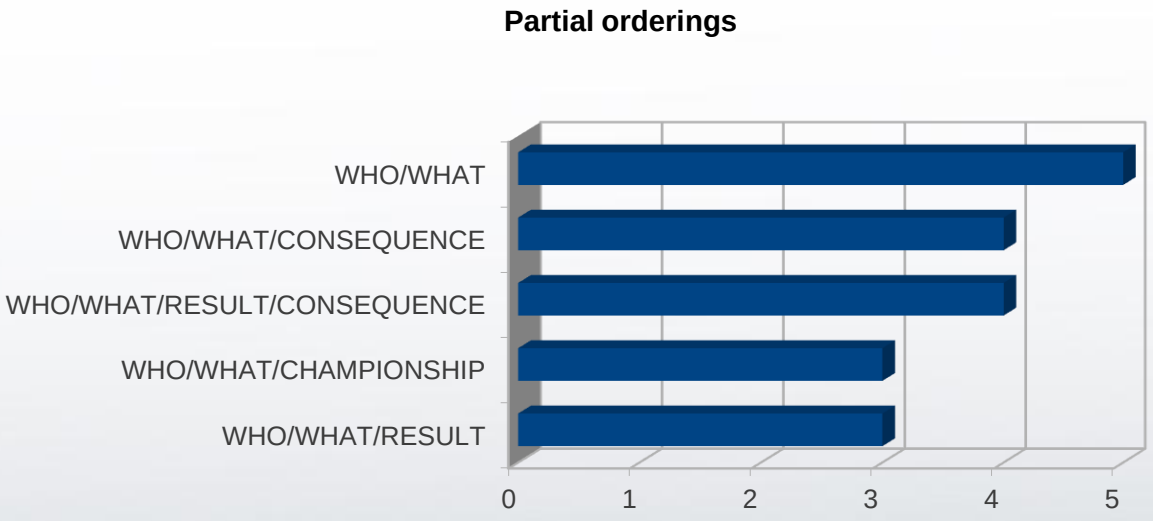
- ***What-e*** is the most frequent aspect
- ***Who***, ***Consequence***, ***How***, ***What***, ***Championship***, and ***Result*** are very frequent and they are present in most summaries

Validation results



- ***Who***, ***consequence***, ***what***, ***championship***, and ***result***
  - Most frequent in 1st paragraph

Validation results



Final remarks

- Specific domain knowledge was necessary for the aspect annotation (at least for the **schedule** aspect)
- Some limitations of our work were the size of corpus of analysis and the number of validation texts
- Future works may be the enrichment of aspects by including ontologies information

Final Remark

- Characterization of human summaries for future works on Multi-Document Summarization

  - It may be possible to suggest new ways of building summaries belonging to “sports” section, for instance:

    - The 1st paragraph ought to contain **who, what, result, championship** and **consequence** aspects, in this order

Thank you for your attention!