PRELIMINARY ASPECTS DISTRIBUTION IN POLITICAL TEXTS

1. Introduction

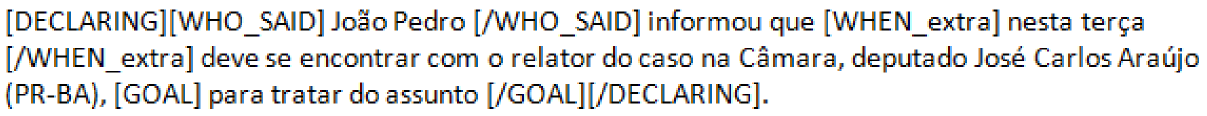
Several attempts to refine Automatic Summarization (AS) models have been pursued, very often through frequency of words, position, or order of text segments (see, e.g., MANI, 2001; JORGE et al., 2011). In multi-document AS, alignment of text spans from diverse sources and redundancy must be also handled. However, they have proven insufficient, and this is evidenced in the Text Analysis Conference (TAC 2010), when novel, aspect-oriented, features for guided summarization were introduced (OWCZARZAK & DANG, 2011; MAKINO et al., 2011). In such track, short and coherent multi-document summaries ought to be produced through pre-defined categories and aspects. Categories mirror subject matter, e.g., Trials, Accidents, or Politics. Their related aspects may help both grasping relevant information from the sources and signaling structuring constraints for related summaries (GENEST et al., 2009). The constraints, in turn, may be described through templates for content organization, resulting in a category-driven process. Numerous studies were developed following such principle. Steinberger et al. (2010) performed deep semantic analyses for modeling aspects in the source documents, for multilingual AS. Makino et al. (2011) compiled aspects from Wikipedia summaries. Barrera et al. (2011) created a question-answering system based on aspects identification for different categories. All of them adopted TAC categories, to know: (1) Accidents and Natural Disasters, (2) Attacks, (3) Health and Safety, (4) Endangered Resources, and (5) Trial and Investigations. Reported regularities for the category-aspects modeling include, e.g., the aspects WHY and DAMAGES for "Accidents and Natural disasters"; WHAT, IMPORTANCE, THREATS, and COUNTERMEASURES for "Endangered Resources" (BARRERA et al, 2011). Some aspects apply more broadly to several categories; others are category-dependent. For example, THREATS is defined under category (1) and targets menaces to natural resources.

Here we report on the replication of TAC guidelines for annotating news texts in Portuguese, extracted from the CSTNews Corpus (http://www.icmc.usp.br/~taspardo/sucinto/cstnews.html) for multi-document AS. Politics was the chosen category. Considering two ways of classifying stories, through the level of cohesiveness or the subject matter of a text (NENKOVA & LOUIS, 2009), we address the latter (OWCZARZAK & DANG, 2011). Our strategy is, thus, meaning-oriented through deeper semantic analyses, and aims at deriving templates for AS. We also investigated if aspects were domain-dependent. For those, category-oriented definitions were provided. We report below on the tagging methodology and derived patterns for multi-document AS.

2. Aspects Tagging Methodology

We annotated only human multi-document summaries of texts on Politics (total of 10), also present in the CSTNews Corpus. Although we adopted the TAC 2010 aspects set, we distinguished general from specific ones. Given our corpus evidences, we redefined these whenever possible and defined new aspects. These were mostly due to category-dependent information. Annotation was first performed individually (three people), and agreement was pursued for tagging consistency.

Aspects hold at two distinct levels: either inter- or intra-sententially, yielding the so-called macro- and micro-tags, respectively. So, concepts identification was pursued through sentence segmentation for the macrostructure, and intra-sentential segmentation, for the microstructure. Aspects for macro-tags are (between brackets, in short for simplicity): ACCUSING (ACCUS*), COMPARING (COMPAR*), CONSEQUENCE (CONSEQ*), DECLARING (DECL), DESCRIBING (DESCR*), INFORMING (INFO), PLEADING (PLEAD*), PREDICTING (PRED), REFERRING (REFER), and REPLYING (REPL). "*" marked tags coincide with the TAC ones. Micro-tags, usually signaling concepts conveyed by single units, comprise GOAL, WHAT; WHEN; WHERE, WHO, and WHY. Except GOAL and WHO, the meanings of the others are about the same as the original ones. WHO was further subdivided in WHO_AFFECTED, WHO_AGENT and WHO_SAID. WHO_AGENT is considered a synonym of PERPETRATOR, originally signaling "Attacks". In Politics, it refers mostly, e.g., to verbal attacks of politicians in political ballots. WHO_SAID signals clear political viewpoints, usually expressed by someone through a declaration (thus, tagged DECL). Figure 1 below shows an example of an excerpt annotated with both, macro- and micro-tags, extracted from our corpus. We borrowed an html-like representation for enclosing text spans that signal the corresponding aspect (*[tag]* .... *[/tag]*).



[DECLARING][WHO_SAID] João Pedro [/WHO_SAID] informou que [WHEN_extra] nesta terça [/WHEN_extra] deve se encontrar com o relator do caso na Câmara, deputado José Carlos Araújo (PR-BA), [GOAL] para tratar do assunto [/GOAL][/DECLARING].

Figure 1. *C42_sumario humano* annotated with aspects

Additionally to aspects tagging, we also produced RST trees (MANN & THOMPSON, 1987) for each summary. Our aim was to draw potential correlations between macro- and micro-tags, and RST relations, in order to substantiate AS templates definitions. Our hypothesis was also that text spans tagged with micro-tags would signal more closely the elementary discourse units in the RST Theory.

3. Results

We produced several statistics from the annotated manual multi-document summaries, aiming at devising prototypical structures and aspects orderings in Politics. Clearly, we could not compute statistical significance due to the small size of our sample. We found that macro-tags may occur either isolated, but still embedding micro-tags, or correlated with other macro- or micro-tags. More than one macro-tag for a text segment signals that there are diverse concept units underlying that segment. Combining macro- and micro-tags is not surprising: distinct levels of representation hold in a discourse, yielding foundation to our distinction between them. Since both macro- and micro-structure hold together for text cohesiveness (van DIJK, 1980; KOCH, 1993), such differing tags can be said to co-occur to assure an adequate thread of information, and thus, coherence (KOCH & TRAVAGLIA, 1995). Figures 2 and 3 show, respectively, the frequency of occurrences (%) of isolated and co-occurring macro-tags in the summary corpus. Their overall frequency is depicted in Figure 4.

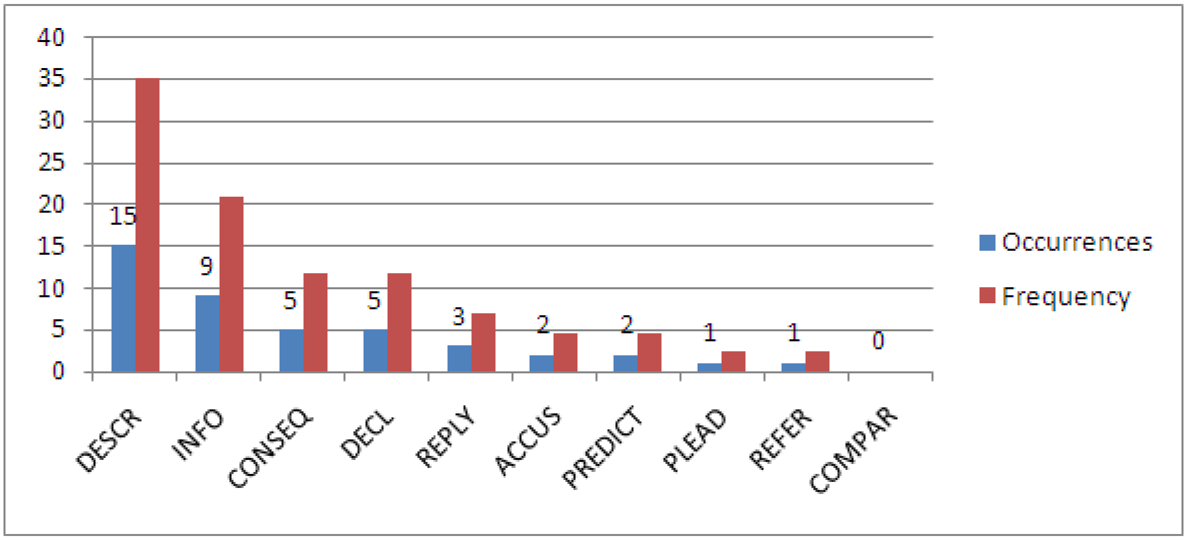


Figure 2. Frequency of isolated macro-tags in the corpus

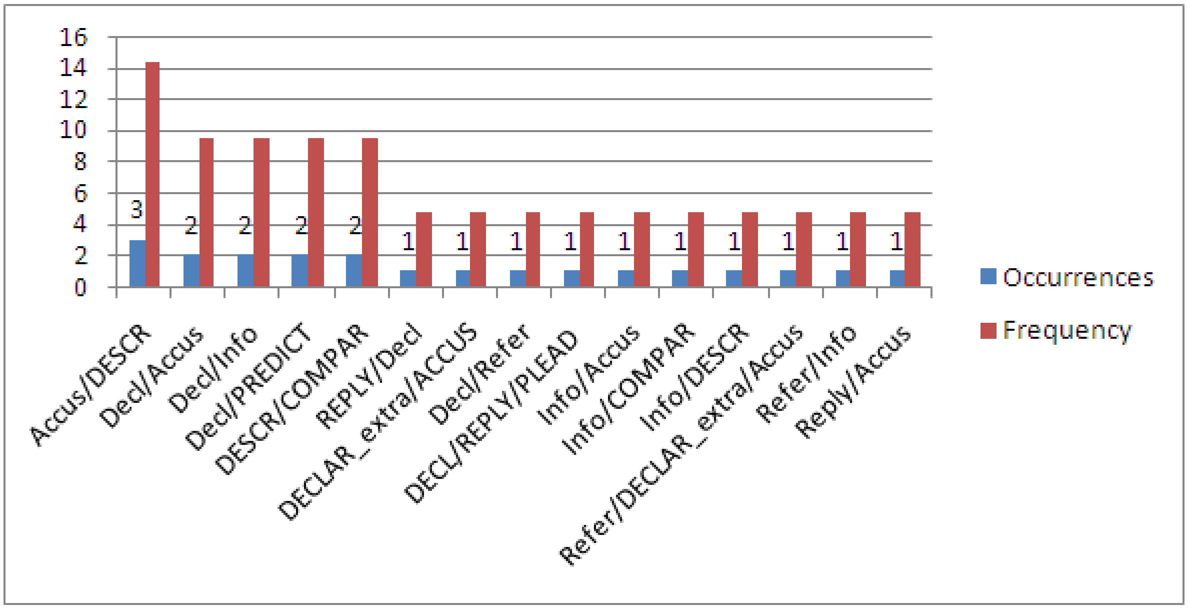


Figure 3. Frequency of co-occurring macro-tags in the corpus

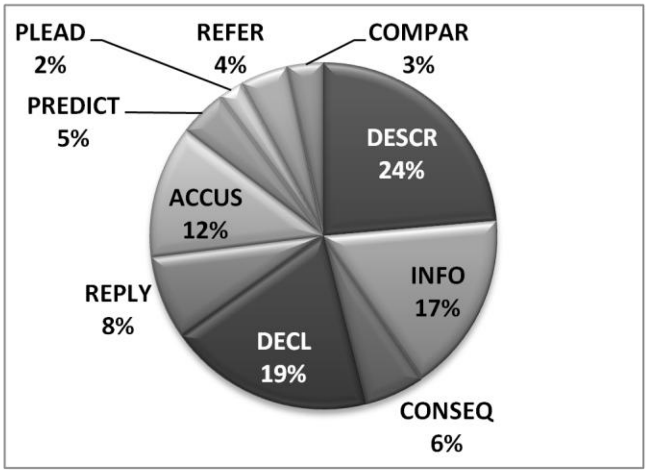
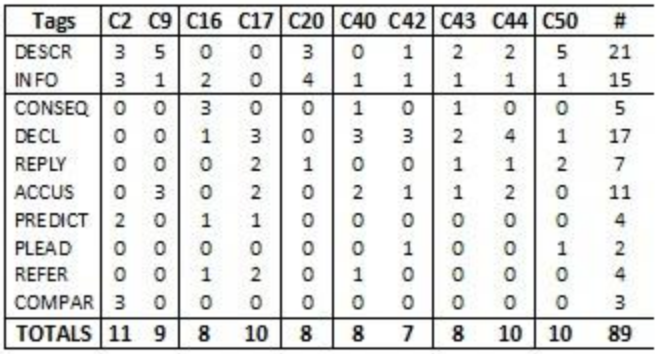


Figure 4. Overall frequency of macro-tags in the corpus

In all, 89 macro-tags were found in the corpus (see Table 1 for their distribution). INFO is the tag chosen for signaling the gist. Occurring 15 times in 10 summaries, it pinpoints more than one gist in some summaries. Actually, summaries C2, C16, and C20 convey such cases. In contrast, C17 does not convey any gist (INFO tag is absent). Actually, the gist is implicit and can be retrieved from C17.

By far the most frequent tag is DESCRIBING, followed by DECLARING, INFORMING, and ACCUSING. DESCR usually signals complementary information to the gist. It conveys data, statistics, examples, or features assigned to political events or people. DECL signals direct or indirect speech, of a politician, in the former case, and of the news article author, in the latter. Its high frequency in our corpus is expected: there is a great incidence of speech or viewpoint turns in Politics texts. Usually, such turns address either a politician referring to another one, or several politicians presenting their opinions on the same issue. ACCUS refers to people mutually accusing themselves. This is quite usual in Brazilian political discourse, as evidenced by c.a. 50% of relative representativeness of such tag in the corpus. It may also refer to law people or government representatives accusing or suing politicians, but this happens less frequently.

Table 1. Overall distribution of macro-tags in the corpus

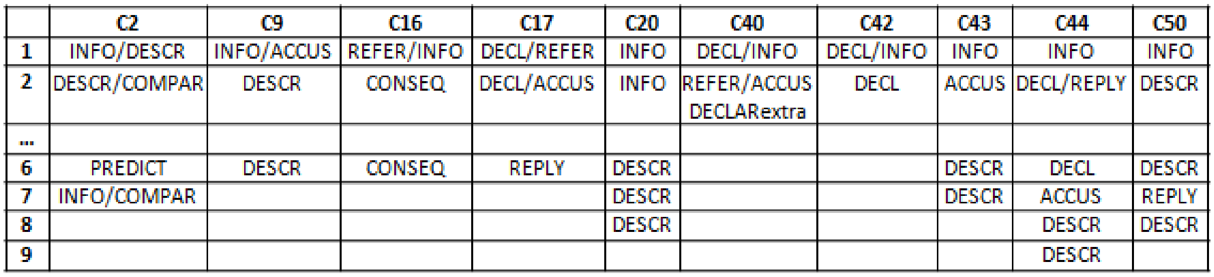


Tags	C2	C9	C16	C17	C20	C40	C42	C43	C44	C50	#
DESCR	3	5	0	0	3	0	1	2	2	5	21
INFO	3	1	2	0	4	1	1	1	1	1	15
CONSEQ	0	0	3	0	0	1	0	1	0	0	5
DECL	0	0	1	3	0	3	3	2	4	1	17
REPLY	0	0	0	2	1	0	0	1	1	2	7
ACCUS	0	3	0	2	0	2	1	1	2	0	11
PREDICT	2	0	1	1	0	0	0	0	0	0	4
PLEAD	0	0	0	0	0	0	1	0	0	1	2
REFER	0	0	1	2	0	1	0	0	0	0	4
COMPAR	3	0	0	0	0	0	0	0	0	0	3
TOTALS	**11**	**9**	**8**	**10**	**8**	**8**	**7**	**8**	**10**	**10**	**89**

Since descriptions are complementary to the main information, and our corpus has been hand-produced by human summarizers, it may be the case that such information has been included in the summaries to comply with the intended compression rate (70%). Declarations, in turn, can yield prototypical patterns for texts on Politics. Actually, if we analyze their co-occurrences (Figure 3), frequent cases of DECL address equally INFO and ACCUS, whilst DESCR and ACCUS co-occur the most in the corpus.

With respect to macro-tag ordering, two patterns were found (see Table 2): the tag INFORMING appears in the 1st sentences of every summary, but C17; and DESCRIBING quite often appears at their endings. Such patterns may just mirror features of the corpus genre (i.e., news): usually the lead (tagged with INFO) is conveyed by sentences at the begin, and additional information (tagged with DESCR) follows.

Table 2. Representativeness of macro-tags in the corpus (per summary)



	C2	C9	C16	C17	C20	C40	C42	C43	C44	C50
1	INFO/DESCR	INFO/ACCUS	REFER/INFO	DECL/REFER	INFO	DECL/INFO	DECL/INFO	INFO	INFO	INFO
2	DESCR/COMPAR	DESCR	CONSEQ	DECL/ACCUS	INFO	REFER/ACCUS DECLARextra	DECL	ACCUS	DECL/REPLY	DESCR
...										
6	PREDICT	DESCR	CONSEQ	REPLY	DESCR			DESCR	DECL	DESCR
7	INFO/COMPAR				DESCR			DESCR	ACCUS	REPLY
8					DESCR				DESCR	DESCR
9									DESCR	

Interesting correlations are also found between macro- and micro-tags. INFORMING co-occurs with WHO_AGENT, WHO_SAID, WHAT, and WHEN, tags that signal the main data referred to in the corresponding political discourses. So, a template might recommend the gist to be conveyed by the 1st sentence, along with information on *who did something, who said something, what who did or said, and when whoever did so.* It should, thus, provide enough information for the reader to grasp what the news is about.

Usually, after the lead sentence, the remaining ones convey extra information. The following two sentences, e.g., often signal DESCR. The others may pinpoint a fact or event (WHAT_extra), someone *who did or said something about the fact or event* (WHO_extra), temporal information (WHEN_extra), none of them related to the gist. Extra information also follows WHY- and GOAL-tagged ones, usually justifying or explaining or indicating the purpose of what has been formerly posed.

Correlating aspects with RST relations was also possible. For example (words in italics indicating RST relations), (i) *Attribution* is always linked to DECL and WHO_SAID; (ii) *Comparison* is usually associated to COMPAR; (iii) *Cause*, *Justify* or *Motivation* may be linked to WHY; (iv) *Purpose* may always be associated to GOAL; (v) *Circumstance* usually is mirrored by WHEN; (vi) *Result* or *Cause* (being volitional or non-volitional) coincides with CONSEQ.

Another interesting outcome of our analyses is that the RST relation *Background* often comes after posing salient information in the news summaries, opposedly to, e.g., scientific texts: *Background* most often comes before that (see, e.g., PARDO, 2005). Such a contrast may be due to genre variations and to writing guidelines: news must present the lead first.

4. Final remarks

Taking after TAC 2010 and 2011 guided summarization track, we explored aspects variations through multi-document summaries of news texts on Politics, aiming at devising AS patterns. We detailed in Section 2 our adopted methodology for aspects tagging, which yielded, besides a consistent procedure for manual tagging, the distinction between aspects that address conceptual information at the micro- and macro-level of each text. We also found through corpus evidences that there is an interplay between macro- and micro-tags, and so there is between macro-tags themselves.

Definitions for both, micro- and macro-aspects, were produced, and may help clarifying and ascertain the validity of such type of annotation. Although the results of applying the presented procedure are still preliminary due to our small sample size, it is certainly scalable for finding writing and information patterns that may be used for mono- or multi-document AS, being this our immediate focus.

We also showed that TAC guidelines are language-free, and this is an evidence for our meaning-oriented approach through deeper semantic analyses. So, we can use them for implementing a wide range of guided-summarization tasks of texts in Portuguese, especially those based on aspect-driven templates.

Moreover, we showed that it is possible to correlate aspects tags with RST relations, and this may be used to enrich AS templates with discourse structures of distinct types.

References

BARRERA, A.; VERMA, R.M.; VICENT, R. SemQuest: University of Houston's Semantics-based Question Answering System. *Proceedings of the Text Analysis Conference*, 2011.

van DIJK, T.A. *Macrostructures: An interdisciplinary study of global structures in discourse, interaction, and cognition*. Hillsdale, N.J.: Lawrence Erlbaum, 1980.

GENEST, Pierre-Etienne; LAPALME, G.; YOUSFI-MONOD, M. HexTac: the Creation of a Manual Extractive Run. *Proceedings of the Text Analysis Conference*, 2009.

JORGE, M.L.R.C.; AGOSTINI, V.; PARDO, T.A.S. Multi-document Summarization Using Complex and Rich Features. In *Anais do VIII Encontro Nacional de Inteligência Artificial*. Natal-RN, 2011.

KOCH, I.V. *A coesão textual*. São Paulo: Contexto, 1993.

KOCH, I.V.; TRAVAGLIA L.C. *A coerência textual*. São Paulo: Contexto, 1995.

LOUIS, A.; NENKOVA, A. Performance confidence estimation for automatic summarization. Proceedings of the 12th *Conference of the European Chapter of the Association for Computational Linguistics*, 541-548, Athens, Greece, 2009.

MAKINO, T.; TAKAMURA, H.; OKUMURA, M. Balanced Coverage of Aspects for Text Summarization. *Proceedings of the Text Analysis Conference*, 2011.

MANI, I. *Automatic Summarization*. John Benjamin's Publishing Company, Amsterdam, 2001.

MANN, W.C.; THOMPSON, S.A. *Rhetorical Structure Theory: a theory of text organization*. ISI/RS-87-190, 1987.

OWCZARZAK, K.; DANG, H.T. Who wrote What Where: analyzing the content of human and automatic summaries. *Proceedings of the Workshop on Automatic Summarization for Different Genres, Media and Languages*, pp. 25-32. Portland, Oregon, June, 23, 2011.

PARDO, T.A.S. *Métodos para Análise Discursiva Automática*. PhD Thesis. Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo. São Carlos-SP, 211p, 2005.

STEINBERGER, J.; TANEV, H.; KABADJOV, M.; STEINBERGER, R. JCR's Participation in the Guided Summarization Task at TAC 2010. *Proceedings of the Text Analysis Conference*, 2010.