CHARACTERIZATION OF TEMPORAL COMPLEMENTARITY: FUNDAMENTALS FOR MULTI-DOCUMENT SUMMARIZATION

Jackson Wilke da Cruz SOUZA[*]
Ariani Di FELIPPO[**]

▪ ABSTRACT: Complementarity is a usual multi-document phenomenon that commonly occurs among news texts about the same event. From a set of sentence pairs (in Portuguese) manually annotated with CST (Cross-Document Structure Theory) relations (Historical background and Follow-up) that make explicit the temporal complementary among the sentences, we identified a potential set of linguistic attributes of such complementary. Using Machine Learning algorithms, we evaluate the capacity of the attributes to discriminate between Historical background and Follow-up. JRip learned a small set of rules with high accuracy. Based on a set of 5 rules, the classifier discriminates the CST relations with 80% of accuracy. According to the rules, the *occurrence of temporal expression in sentence 2* is the most discriminative feature in the task. As a contribution, the JRip classifier can improve the performance of the CST-discourse parsers for Brazilian Portuguese

▪ KEYWORDS: Linguistic description. Complementarity. CST. Multi-document Summarization. Natural Language Processing.

Introduction

The access and availability of digital information have been grown very rapidly. According to the projections of Taufer (2013), digital information will amount to 44 zettabytes by 2020. Some Natural Language Processing (NLP) subareas seek to develop computational applications capable of handling this tremendous amount of data.

One of these subareas is Multi-Document Summarization (MDS), which goal is to automate the production of a summary given a group of texts on the same topic compiled from different sources (MANI, 2001). The vast majority of the MDS work relies on the production of extractive summaries (or extracts), which are composed of sentences taken exactly as they appear in the source texts. Such summaries tend to

[*] Federal University of São Carlos (UFSCar) / Interinstitutional Nucleus of Computational Linguistics (NILC), São Carlos - São Paulo - Brasil. jackcruzsouza@gmail.com

[**] Federal University of São Carlos (UFSCar) / Interinstitutional Nucleus of Computational Linguistics (NILC), São Carlos - São Paulo - Brasil. Department of Letters (DL). arianidf@gmail.com

be *informative*, since they convey the central content of the collection to the point of replacing the reading of the source texts, and *generic,* i.e., they are aimed at a broad readership community (KUMAR; SALIM; RAZA, 2012).

The multi-document summaries have been generated in three stages: (i) *analysis* (i.e., interpretation of the source collection to obtain an internal representation of its content); (ii) *transformation* (i.e., the main stage of the summarization model; it takes the internal representation of the source texts to produce the summary internal representation), and (iii) *synthesis* (i.e., the summary internal representation is linguistically realized into the final summary) (SPARCK-JONES, 1993; MANI, 2001).

According to the amount and level of linguistic knowledge, MDS can be shallow or deep (MANI, 2001). A shallow MDS method uses little or no linguistic knowledge; it usually performs a statistical analysis of source texts in order to produce extracts. Shallow methods/systems usually require low-cost processing and achieve higher robustness and scalability. However, the extracts generated by these methods tend to suffer from lack of text coherence and cohesion, and from low informativeness. Deep MDS approaches, in turn, use sophisticated linguistic knowledge codified in grammars, semantic repositories and discourse models, and thus they have a restricted application and a high cost of development. On the other hand, the deep MDS paradigm generates extracts with less linguistic problems and also *abstracts*.

In order to produce informative and generic extracts, it is necessary to select the most important sentences of the collections, avoiding redundancy and contradiction, and keeping complementarity among them. Redundancy, contradiction and complementarity are some of the so-called multi-document phenomena whose identification and treatment are important for the linguistic quality and informativeness of the multi-document extracts. We illustrate the three phenomena with the sentence pairs (S1 and S2) in (1), (2), and (3) (with the original Portuguese sentences in parenthesis).

(1) *Redundancy*
S1: The margin of error is plus or minus 2 percentage points. (*A margem de erro é de dois pontos percentuais, para mais ou para menos.*)
S2: The margin of error is 2 percentage points. (*A margem de erro é de 2 pontos porcentuais.*)
(2) *Complementarity*
S1: For Japan, the reported magnitude 6.8 is considered "strong". *(No caso do Japão, a magnitude apontada de 6,8 é considerada "forte".)*
S2: In Niigata, after an earthquake of same magnitude (6.8) in October 2004, 65 people were killed and more than 3,000 injured. *(Em Niigata, um terremoto em outubro de 2004, também de magnitude 6,8, matou 65 pessoas e deixou mais de 3.000 feridos.)*
(3) *Contradiction*
S1: Both José Maria Eymael (PSDC - Christian Social Democratic Party) and Rui Pimenta (PCO - Workers' Cause Party) have not reached 1% of voting

intentions. *(José Maria Eymael, do PSDC, e Rui Pimenta, do PCO, não chegaram a obter 1% das intenções de voto.*)

S2: Candidates José Maria Eymael (PSDC) and Ruy Pimenta (PCO) did not score. *(Os candidatos José Maria Eymael (PSDC) e Ruy Pimenta (PCO) não pontuaram.)*

The sentences in Example (1) are redundant because the main content is very similar. In Example (2), the sentences are complementary, since they share some information (*6.8 magnitude earthquakes in Japan*), but S2 provides additional information not presented in S1. Specifically, S1 provides historical information on *a 6.8 magnitude earthquake that struck a region of Japan in 2004*. Finally, in Example (3), S1 and S2 contradict each other because, in S2, the candidates did not receive any vote, and, in S1, they received some votes, but the number of votes did not add up to 1% of the intentions.

Specifically, the identification of such phenomena during the texts analysis is very important because: (i) the most redundant sentences of the collection convey its main content and thus must be included in the summary; (ii) relevant and complementary sentences should also be included in the summary, and (iii) redundant or contradictory information should not be selected for the summary. In order to do that, linguistic descriptions of the multi-document phenomena are essential, since they provide clues to be followed by the MDS methods. In this paper, we focus on the (temporal) complementarity, since redundancy (e.g., HATZIVASSILOGLOU et al., 2001; NEWMAN et al., 2004; HENDRICKX et al., 2009; SOUZA; DI-FELIPPO; PARDO, 2013) and contradiction (e.g., CONDORAVDI et al., 2003; MARNEFFE; RAFFERTY; MANNING, 2008; MARNEFFE, 2012) are the phenomena that have been investigated more extensively in the literature.

In Section 2, we describe the multi-document relations of the CST (*Cross-Document Structure Theory)* model (RADEV; JING; BUDZIKOWSKA, 2000) that codify complementarity and the main methods of identifying the CST relations. In Section 3, we present the *corpus* used in this work and then the selection of temporal complementary pairs of sentences from the *corpus*. Section 4 describes the linguistic characteristics of the temporal complementarity and the translation of such characteristics into attributes or features that can be used to automatically identify the CST relations. In Section 5, we describe the linguistic description of the *corpus* that is necessary to evaluate the potential of the attributes to detect the temporal complementarity. Finally, in Section 6, we present the evaluation results regarding the potential of the attributes to detect the CST relations of temporal complementarity and some final remarks.

Related studies

Two sentences from different texts on the same topic can be related to each other in a number of ways (MANI, 2001). The analysis of the relationships among such sentences (i.e., multi-document or intertextual analysis) has been extensively investigated during the last decades in the NLP field. An example of NLP application that benefits from such analysis is Multi-Document Summarization (MDS), which aims at generating a unique summary from the content of several source texts. The investigation of cross-document relationships provided a set of rhetorical relations identified among sentences from topically related documents. These rhetorical relations are based on the CST model (RADEV; JING; BUDZIKOWSKA, 2000).

The CST model allows the connection (in pairs) of semantically related textual units (e.g., sentences) from documents on the same topic. Originally, it was proposed a set of 24 CST relations (Table 1).

Table 1 – Original set of CST relations.

Identity	*Modality*	*Judgment*
Equivalence	*Attribution*	*Fulfillment*
Translation	*Summary*	*Description*
Subsumption	*Follow-up*	*Reader profile*
Contradiction	*Elaboration*	*Contrast*
Historical background	*Indirect speech*	*Parallel*
Cross-reference	*Refinement*	*Generalization*
Citation	*Agreement*	*Change of perspective*

Source: Radev (2000).

Some authors have refined the original CST relations, proposing more compact sets (e.g., ZHANG; OTTERBACHER; REDEV, 2003; MAZIERO, 2012; MAZIERO; JORGE; PARDO*,* 2014). For Brazilian Portuguese, the original set was reduced to 14 relations and they were organized into two groups (MAZIERO, 2012; MAZIERO; JORGE; PARDO*,* 2014): (i) content relations (Identity, Elaboration, Equivalence, Contradiction, Summary, Subsumption, Overlap, Historical background and Follow-up), and (ii) form relations (Attribution, Citation, Modality, Indirect Speech and Translation). The content relations, in particular, codify the following multi-document phenomena: redundancy, complementarity and contradiction.

A number of papers have addressed the benefits of CST for MDS. The study proposed by Zhang, Blair-Goldensohn and Radev et al. (2002) was the first to consider multi-document structural relationships, codified by the CST relations, to generate a summary. Specifically, the authors first use MEAD (RADEV; JING; BUDZIKOWSKA, 2000; RADEV et al., 2003), a summarization system based on *cluster centroids*[1], to rank the source sentences and produce an initial extract. Then the low-salience

[1] In general, the analysis of the source texts in cluster and centroid-based MDS methods consists in grouping sentences that are highly similar to each other. Thus, the clusters with similar sentences represent the “topics” of the collection. Each cluster is represented by a centroid, i.e., a set of statistically significant words. The cluster centroids are used to

sentences ranked by MEAD are replaced by sentences that have more CST relations in the collection, which tend to be more informative.

Jorge and Pardo (2010), focusing on Brazilian Portuguese, also apply CST in MDS. In this work, they rank sentences according to the number of CST relations they have in the collection. More recently, Cardoso (2014), also using a *corpus* in Brazilian Portuguese, developed a MDS method that integrates CST, Rethorical Structure Theory (RST) (MANN; THOMPSON, 1987), and subtopics to model the summarization process. All the mentioned research works deal with manually annotated *corpora*, but a CST annotation is an expensive and time-consuming task, since it requires a highly trained team of experts capable of producing a considerable amount of data.

In order to solve this problem, there have been efforts put forth to automatically identify CST relations in texts using Machine Learning (ML)[2] techniques. Zhang et al. (2003) focused on the detection of six CST relations across source sentences. The developed classifier[3] was able to efficiently identify unrelated sentence pairs, but showed poor performance in classifying the type of relations.

Miyabe, Takamura and Okomura (2008) attempted to detect the Equivalence and Transition relations. The method of the authors to automatically identify these relations necessarily starts with the detection of Equivalence and then Transition (i.e., relation that occurs between sentences with the same information, differing by numerical values; this relation would be similar to the CST Contradiction relation).

In Zahri and Fukumoto (2011), the identification of Identity, Paraphrase (similar relation to CST Equivalence), Subsumption, Overlap, and Elaboration is a stage of a summarization application. In such MDS system, the authors used the headlines of the documents to extract sentences with salient terms from the source texts using the statistical model. Then they assigned rhetorical relations among those sentences that were learned by a ML algorithm. Finally, they ranked the sentences by measuring their relative importance within the source collection through the method called PageRank[4] (ERKAN et al*.,* 2004), and selected the most salient sentences to compose the extract. According to the authors, the combination of PageRank along with rhetorical relations among sentences helps to avoid the generation of extracts with redundant information.

Kumar, Salim and Raza (2012) applied *linguistic* (e.g., verbal similarity between two sentences), *structural* (e.g., length sentence), and *statistic* (e.g., word overlap) features

identify the sentences in each cluster that are most similar to the centroid. Thus, the system selects the sentence that is most relevant to each cluster. For more details on cluster and centroid, see Jurafsky and Martin (2009).

[2] Machine learning is a field of Artificial Intelligence that aims at exploring the study and construction of algorithms that can learn from and make predictions on data.

[3] In Machine Learning, a classifier is an algorithm that takes a set of parameters (or features) that characterize objects (or instances) and uses them to determine the type (or class) of each object. Using a training set containing a list of instances with known classifications, since each class is described by a set of attributes or features, the classifier decides how the parameters ought to be weighed and combined in order to separate the various classes of instances. Then, the weights determined in the training set are applied to a set of instances that do not have known classes in order to determine what their classes are likely to be (MITCHELL, 1997).

[4] PageRank is an algorithm used by Google Search to rank websites in their search engine results (https://en.wikipedia.org/wiki/PageRank).

for the automatic identification of four CST relations: Identity, Overlap, Subsumption, and Description. To evaluate the method, the authors used a dataset taken from CSTBank (RADEV; OTTERBACHER; ZHANG, 2004), a multi-document *corpus* of English news articles whose sentences were annotated with CST relationships. Specifically, Kumar et al. selected 476 sentence pairs for training and 206 sentence pairs for testing. In order to identify the CST relations, they used three different ML techniques, and the results showed that Identity is the easiest relation to be detected (f-measure > 90%). The authors point out that this result may be related to the high lexical similarity between the sentences linked by Identity, which facilitates the automatic identification of the relation.

For Brazilian Portuguese, there are a few studies on automatic detection of CST relations (MAZIERO, 2012; MAZIERO; JORGE; PARDO, 2014; SOUZA; DI-FELIPPO; PARDO, 2012, 2013).

Maziero (2012) and Maziero, Jorge and Pardo (2014) developed the multi-document discourse parser called CSTParser, which applies the most popular set of similarity features to detect a CST relation between two sentences: (i) difference in number of words between the sentences (S1–S2); (ii) percentage of words in common between S1 and S2; (iii) position of S1 in the text (0 – beginning: first three sentences; 1 – middle; 2 – end: last three sentences); (iv) position of S2 in text (the same as above); (v) number of words in the longest common substring between S1 and S2; (vi) difference in the number of nouns between S1 and S2 (common and proper nouns); (vii) difference in the number of adverbs between S1 and S2; (viii) difference in the number of adjectives between S1 and S2; (ix) difference in the number of verbs between S1 and S2; (x) difference in the number of proper nouns between S1 and S2; (xi) difference in the number of numerals between S1 and S2, and (xii) number of possible synonyms in common in S1 and S2. The CSTParser achieved an overall accuracy of 68,13%, which is the average accuracy obtained by a classifier for Overlap, Subsumption, Elaboration, Equivalence, Historical background and Follow-up, and by hand-crafted rules for Identity, (explicit) Contradiction, Attribution, Indirect Speech, and Translation[5]. According to the authors, this overall accuracy is a good result given the subjectivity of the cross-document analysis.

Souza, Di-Felippo and Pardo (2012, 2013) focused on the automatic detection of the CST relations that represent redundancy (Identity, Equivalence, Summary, Subsumption and Overlap) and the types of redundancy (i.e., total, partial, or null) codified by these relations (2012). To investigate the formal characteristics of redundancy, the authors used CSTNews (CARDOSO et al., 2011), a multi-document *corpus* of journalistic texts that were annotated at sentence level with CST relations. Besides sentence position in source texts, the authors applied the following attributes: (i) word overlap,[6] (ii) noun and verb overlap, (iii) morphosyntactic pattern (e.g., noun+preposition+noun) overlap,

[5] Summary, Modality and Citation were not included in the method proposed by Maziero (2012) due to the low frequency in the training *corpus*.

[6] Souza, Di-Felippo and Pardo (2012, 2013) used a traditional *word overlap* measure to detect lexical similarity between sentences. The measure calculates the number of words in common between two sentences. The authors also applied

(iv) subject overlap (i.e., occurrence or identical subjects) (v) main verb overlap, (v) object (direct or indirect) overlap, (vii) part-of-speech tag overlap, and (viii) occurrence of synonyms. Using PART (WITTEN; FRANK, 1998) and J48 (QUINLAN, 1993) ML algorithms, Souza, Di-Felippo and Pardo (2013) showed that the classifier based on all attributes identifies the redundancy types (total, partial and null) with 97.7% of accuracy, and the CST relations with 62.2% of accuracy. The second best classifier uses only one attribute (noun overlap) and achieves 91.1% of accuracy for the types and 60% for the CST relations.

Based on this review of the literature, we observed that several works have been investigated the automatic detection of CST relations, especially those relations that represent redundancy. However, the CST relations that refer to complementarity (i.e., Follow-up, Historical background and Elaboration) have been automatically detected using attributes that characterize redundancy, such as those investigated by Souza, Di-Felippo and Pardo (2012, 2013), which are not specific of complementarity.

CSTNews *corpus* and complementarity

In order to describe the complementarity phenomenon, we used CSTNews, a reference multi-document *corpus* in Brazilian Portuguese (CARDOSO et al., 2011). The CSTNews *corpus* is composed of 50 clusters (or collections), totaling 140 source texts, 2,088 sentences and 47,240 words. The clusters are organized into the following categories: *world* (14), *politics* (11), *daily news* (13), *science* (1), *money* (1), and *sports* (10). The source texts were compiled from the following online news agencies *Folha de São Paulo, Estadão, O Globo, Jornal do Brasil,* and *Gazeta do Povo*.

Each cluster is composed of: (i) two (2) to three (3) sources texts; (ii) mono-document summaries; (iii) six (6) multi-document abstracts and six (6) manual multi-document extracts; (iv) one (1) automatic multi-document extract; (v) manual CST annotation among source texts; (vi) manual annotation of temporal expression of each source texts; (vii) part-of-speech tagging and syntactic parsing of each source text; (viii) semantic annotation of nouns and verbs with their corresponding Princeton WordNet[7] synsets (FELLBAUM, 1998); (ix) manual annotation of aspects (e.g., *who, what, where, when,* etc.) of one (1) source text; (x) annotation of each source document with RST[8] relations (MANN; THOMPSON, 1987), and (xi) manual subtopic segmentation of each source text.

two variations of *word overlap*, i.e., *noun overlap* and *verb overlap*, to compute similarity based on the number of nouns and verbs shared by two sentences.

[7] WordNet is an on-line lexical reference system whose design is inspired by current psycholinguistic theories of human lexical memory. English nouns, verbs, adjectives, and adverbs are organized into synonym sets (or synsets) (e.g., {car, auto, automobile, machine, motorcar}), each representing one underlying lexical concept. Different relations link the synonym sets, such as antonymy, hyponymy/hypernymy, meronymy/holonymy, entailment, and cause. More details on Princeton WordNet are available at <https://wordnet.princeton.edu/>.

[8] The RST model allows the analysis of text coherence. Such analysis consists of verifying if the elementary discourse units (EDUs), which are the minimal building blocks of a discourse structure, are interconnected. Each EDU in a

Based on the CST typology proposed by Maziero et al. (2012), we selected a set of sentence pairs from CSTNews annotated with the following CST relations: Follow-up, Historical background and Elaboration. In order to select the pairs, we used the CSTNews online interface[9]. According to Maziero et al. (2012), Historical background and Follow-up are content relations that codify two different types of temporal complementarity, as illustrated in examples (i) and (ii) of Table 2, respectively.

Table 2 – Examples of temporal complementarity.

Temporal complementarity	**Sentence pair**
(i) S2 presents some historical background about the event described in S1 (S1←S2); the historical event is the focus of S2 (*Historical background*)	S1: A plane crash in Bukavu, a city in eastern Democratic Republic of the Congo (DRC) killed 17 people on Thursday, said a United Nations spokesman on Friday. (*Um* ***acidente aéreo na localidade de Bukavu, no leste da República Democrática do Congo (RDC)****, matou 17 pessoas na quinta-feira, informou nesta sexta-feira um porta-voz das Nações Unidas*.) S2: Air accidents are frequent in Congo, where 51 private companies operate oldplanes built in the former Soviet Union. (***Acidentes aéreos*** *são frequentes* ***no Congo****, onde 51 companhias privadas operam com aviões antigos principalmente fabricados na antiga União Soviética*.)
(ii) S2 presents additional information which has happened since S1 (S1←S2); the events in S1 and S2 are related and have a relatively short period of time between them (*Follow-up*)	S1: **The secondary runway at São Paulo's Congonhas Airport opened** at 6 am, only for departures. (*A pista auxiliar de Congonhas abriu às 6h, apenas para decolagens*.) S2: **São Paulo's Congonhas Airport only opened** for landings at 8:50 am. (*Congonhas só abriu para pousos às 8h50.*

Source: Own elaboration.

relation is classified as nuclei (i.e., more important propositions) or satellites (i.e., complementary information). When coherent, the units of a text are linked to each other by rhetorical relations (also known as coherence or discourse relations). Relations with one nucleus and one satellite are said to be mononuclear relations. When all the interconnected units are nuclei, we have a multinuclear relation.

[9] <http://nilc.icmc.usp.br/CSTNews/>.

We illustrate the complementarity in (i) with a pair of sentences from distinct news reporting a plane crash in Congo. Specifically, the sentences present common content (in bold), and S2 provides historical information (underlined) about the other, which is the frequent occurrence of plane crashes in Congo (since the commercial carriers operate with old aircrafts). According to the typology of Maziero et al. (2012), Historical Background is the CST relation that represents the complementarity illustrated in (i), characterizing as temporal; in this case, the temporal complementarity is related to the frequent aspect of the main event described in S1.

To illustrate the complementarity in (ii), we provide a pair of sentences compiled from distinct news reporting delays and cancellations of flights at Congonhas Airport due to bad weather. The temporal complementarity relation between S1 and S2 occurs because they share some information (*time which Congonhas Airport opened*), and the event reported in S2 followed the event described in S1 after a short period. According to the typology of Maziero, Jorge e Prado (2014), this type of temporal complementarity is codified by Follow-up.

Elaboration, unlike the relations illustrated in Table 2, does not involve temporal information. The example in Table 3 illustrates this non-temporal complementarity.

Table 3 – Example of non-temporal complementarity.

NON-TEMPORAL COMPLEMENTARITY	SENTENCE PAIR
S2 details some information present in S1 (S1←S2); S2 does not repeat information present in S1 and the additional information present in S2 is the focus of S2. (*Elaboration*)	S1: Although the improvement project is approved, the **work schedule** has not been released yet. (*Apesar da definição, o cronograma da obra não foi divulgado*.) S2: The **work schedule** depends on final studies being carried out by Infraero. *(O cronograma da obra depende de estudos finais que estão sendo realizados pela Infraero*

Source: Own elaboration.

The sentences in Table 3 were extracted from news reporting a renovation project of the Congonhas Airport. We observed that S1 and S2 share some content (the project schedule), and S2 details some information present in S1. The additional information in S2 is the focus of the sentence and consists of the reason why the project timeline has not been published (it depends on final studies being carried out by Infraero – Brazil's national airport authority).

Table 4 shows the total number of sentence pairs annotated with the CST relations that express the different types of complementarity in CSTNews.

Table 4 – The statistics of complementary in the CSTNews *corpus*.

Complementarity	CST relation	Qt. of pairs	Total
Non-temporal	Elaboration	343	343
Temporal	Historical background	77	370
	Follow-up	293	
			713

Source: Own elaboration.

In Table 4, we see that non-temporal complementarity, expressed by Elaboration, occurs in 343 sentence pairs. We also see that temporal complementarity occurs in 370 sentence pairs, with 293 cases of Follow-up and 77 cases of Historical background. Thus, there are 713 sentence pairs with complementarity in CSTNews.

To date, we manually analyzed 45 pairs of each CST relations that codify temporal complementarity, totaling a *subcorpus* with 90 pairs. The manual analysis of the *subcorpus* allowed us to detect linguistic characteristics of temporal complementarity, represented by Historical background and Follow-up. In the next section, we describe these linguistic characteristics and present their conversion into machine-tractable attributes or features.

Identifying machine tractable attributes

Two complementary sentences extracted from different news texts reporting the same event are relatively similar, which can be seen in the examples that illustrate the complementarity phenomenon. Such similarity underlies the very definition of the CST relations, since Zhang and Radev (2005) state that such relations only occur between semantically related sentences. Based on the degree of similarity, the classes of content CST relations in the Maziero, Jorge e Prado (2014) typology can be organized into the following hierarchy: redundancy > complementarity > contradiction. Thus, complementarity is a multi-document phenomenon that involves intermediate level of similarity. It is not known, however, if the different temporal complementarities, expressed by Historical background and Follow-up CST relations, present the same level of similarity. Thus, we decided to investigate the potential of redundancy as a distinctive feature. In order to make it possible, we have selected the most efficient attributes that have been used in the literature to automatically detect similarity among sentences.

According to Hatzivassiloglou et al. (2001), Newman et al. (2004) and Souza, Di-Felippo and Pardo (2012, 2013), there are three efficient features to capture similarity or redundancy between two sentences, which are: (i) noun overlap, (ii) sentence position (in the source text), and (iii) subtopic overlap. Noun overlap is an efficient measure since most of the frequent words that are content words in texts or sentences are nouns. As we mentioned, *noun overlap* (Nol) is a version of the traditional measure *word overlap.* Calculation of Nol is shown in (4). The result value of *noun overlap* varies between zero and one. If the value is one or close to one, it means high similarity, but, if the value is zero or close to zero, it means the opposite, i.e., low similarity between sentences.

$$\text{(4)}\quad \text{Nol (S1, S2)} = \frac{\text{\# Common nouns (S1,S2)}}{\text{\# Nouns (S1) + \# Nouns (S2)}}$$

Concerning (ii), Souza Di-Felippo and Pardo (2012, 2013) state that sentence position indicates similarity because news texts commonly follow the inverted-pyramid structure. Such structure illustrates how information is organized in the following blocks of decreasing relevance: (i) lead, i.e., the main information of the news text; it corresponds to the first paragraph; (ii) body, i.e., paragraph(s) that follow(s) the lead, illustrating the main information of the text, and (iii) closer, i.e., the final paragraph which reinforces or debunks the lead (LAGE, 2002).

Thus, if two sentences occupy close positions in their correspondent source texts, they tend to convey similar information. If the distance between the positions they occupy in the texts is long, the content of the sentences tend to be not so similar. Under this hypothesis, Souza, Di-Felippo and Pardo (2012) proposed the *sentence distance* attribute, described in (5). For example, given a sentence pair from a cluster *x* (S1 and S2), where S1 is the Sentence 6 from the Text 1 and S2 is the Sentence 4 from the Text 2, the distance value between them is equal to 2 (positions). In this paper, the distance value is divided by the longest distance between two sentences identified in the *subcorpus*, which normalizes the feature by document length. Thus, the equation described in (5) generates a distance value between zero and one. If the value is zero or close to zero, it means high similarity between sentences; if the value is one or close to one, it means the opposite, i.e., low similarity.

$$\text{(5)}\quad \text{Distance (S1, S2)} = \frac{\text{\# Distance between S1 and S2}}{\text{\# Longest distance in } \textit{subcorpus}}$$

The attribute (iii), *subtopic overlap*, is a refinement of the sentence distance feature. In the inverted pyramid structure, the lead is the main information (or topic) of a piece of news and the details about the lead are the subtopics, which are directly or indirectly linked to the topic according to thematic progression (KOCH, 2009). Due to the inverted pyramid structure, sentence position indicates content similarity. However, the inverted pyramid is just a writing guideline and thus a piece of news might present relatively different structure. Consequently, it is not always possible to capture redundancy based on sentence position. Thus, subtopic overlap might be a relevant feature to identity redundancy since it is independent of the position occupied by the sentences. Moreover, once *subtopic overlap* is a semantic feature, we say that it is richer or more informative than noun overlap, which is based on word forms. In order to capture similarity based on subtopic, we proposed the *subtopic overlap* attribute (SubTol). As a binary feature, the SubTol values can be "yes" or "no", indicating whether a sentence pair conveys the same subtopic or not.

Besides similarity or redundancy, temporal complementarity, expressed by the Historical background and Follow-up CST relations, involves the occurrence of temporal marks, such as simple adverbs and expressions.

The Example (6a) illustrates Historical background. In (6a), we observe the occurrence of temporal expressions only in S1, i.e., *on Thursday* (*na quinta-feira*) and *on Friday* (*nesta sexta*). In Example (6b), the two temporal expressions *on Thursday night* (*na noite desta quinta-feira*) and *since last 13th (July)* (*desde o último dia 13*) occur in S1, and the only expression *in 1996* (*em 1996*) occurs in S2.

(6a) S1: A plane crash in Bukavu, in the Eastern Democratic Republic of the Congo (DRC) killed 17 people on Thursday, said a United Nations spokesman on Friday. (*Um acidente aéreo na localidade de Bukavu, no leste da República Democrática do Congo (RDC), matou 17 pessoas na quinta-feira, informou nesta sexta-feira um porta-voz das Nações Unidas*.)
S2: Air accidents are frequent in Congo, where 51 private companies operate elderly planes built in the former Soviet Union. (*Acidentes aéreos são frequentes no Congo, onde 51 companhias privadas operam com aviões antigos principalmente fabricados na antiga União Soviética*.)

(6b) S1: TAM airlines confirmed on Thursday night that the right thrust reverser was deactivated since last 13th (July). (*A TAM confirmou, na noite desta quinta-feira, que o Airbus da TAM estava com o reverso do lado direito desligado, desde o último dia 13*.) S2: In 1996, a failure in the reverser caused an accident with a TAM Fokker-100, occurred seconds after takeoff from Congonhas Airport. (*Em 1996, uma falha no reverso foi a causa do acidente com o Fokker-100 da TAM, ocorrido segundos depois da decolagem, também em Congonhas*.)

The sentence pair in Example (7a) illustrates the Follow-up relation. Follow-up involves the occurrence of temporal expression in both sentences of the pair. The expression *at 8:50 a.m.* (às 8h50) in S2 indicates the complementary event, which occurred after the event described in S1 (*the secondary runway opened at 6 am)*. In Example (7b), the temporal *during this Sunday, day 6* (*durante este domingo, dia 6*) occur in S1 and the adverb *after* (*depois*) in S2.

(7a) s1: The secondary runway at São Paulo’s Congonhas Airport opened at 6 am only for departures. (*A pista auxiliar de Congonhas abriu às 6h, apenas para decolagens*.) S2: São Paulo’s Congonhas Airport opened for landings only at 8:50 am. (*Congonhas só abriu para pousos às 8h50*.)

(7b) S1: During this Sunday, day 6, bloody fights occurred (*Durante este domingo, dia 6, foram travadas lutas sangrentas*.)

S2: The Israeli offensive was launched after a string of Hezbollah attacks on Sunday that caused the biggest casualties for Israel in the four weeks of the conflict. (*A ofensiva israelense foi lançada depois de uma sequência de ataques do Hezbollah no domingo que causou as maiores baixas para Israel nas quatro semanas do conflito*.)

Thus, to use these linguistic clues for the automatic detection of temporal complementarity, we proposed four (4) binary attributes: occurrence of temporal expression in S1 (TES1), occurrence of ET in S2 (TES2), occurrence of adverb in S1 (ADVS1) and occurrence of adverb in S2 (ADVS2).

Table 5 synthesizes the total set of seven (7) attributes used in this research.

Table 5 – Temporal complementarity atributes.

Attribute	Description	Acronym
Noun overlap	It captures redundancy based on the number of nouns the sentences of a pair have in common.	Nol
Distance	It captures redundancy based on the distance between the positions/locations occupied by the sentences of a pair in their correspondent source text	Distance
Subtopic overlap	It captures redundancy based on the subtopic overlap between the sentences of a pair.	SubTol
Occurrence of temporal expression in S1	It captures temporal complementarity by the occurrence of temporal expression in Sentence 1 of a pair	TES1
Occurrence of temporal expression in S2	It captures temporal complementarity by the occurrence of temporal expression in Sentence 2 of a pair	TES2
Occurrence of adverb in S1	It captures temporal complementarity by the occurrence of adverbs in Sentence 1 of a pair	ADVS1
Occurrence of adverb in S2	It captures temporal complementarity by the occurrence of adverb in Sentence 1 of a pair	ADVS2

Source: Own elaboration.

Next, we describe the linguistic characterization of the 90 sentence pairs to evaluate the potential of the attributes.

Linguistic description of the *corpus*

The characterization process of the 90 sentence pairs for the automatic evaluation of the features consisted in describing the sentences in a way that we could calculate the seven features or attributes for each pair. Specifically, the sentence description was based on previous annotations of the CSTNews *corpus*.

In order to calculate NoI, Distance, ADVS1 and ADVS2 attributes, we identified the nouns and adverbs that constitute each sentence and the position of the sentences in their source texts. The nouns and adverbs and also the sentence position were compiled from the syntactic annotation of CSTNews, which was automatically performed by the parser[10] PALAVRAS (BICK, 2000). We manually revised the automatic annotation of CSTNews, in order to reduce *noise*[11] (i.e., tagging errors) and *silence* (i.e., tagging omissions) produced by the parser. Table 6 illustrates the syntactic annotation[12] of the sentence *São Paulo's Congonhas Airport only opened for landings at 8:50 am* (*Congonhas só abriu para pouso às 8h50*). The annotation in Table 6 includes the position the sentence ("s4") and the following nouns, *Congonhas* (*pos*[13]*=np*") and *pousos* (landings) (*pos*="*n*"). For the linguistic description, we selected the values of the attribute *lemma* of nouns and adverbs. For describing the sentence in Table 6, for example, we selected *pouso* (landing) and *Congonhas.*

Table 6 – Example of syntactic annotation of CSTNews.

```
</s><s id=”s4” text=”Congonhas só abriu para pousos, às 8h50.”>
    <terminals>
        <t id=”1” word=”Congonhas” lemma=”Congonhas” pos=”np”/>
        <t id=”2” word=”só” lemma=”só” pos=”adv”/>
        <t id=”3” word=”abriu” lemma=”abrir” pos=”v-fin” morph=”PS 3S IND VFIN”/>
        <t id=”4” word=”para” lemma=”para” pos=”prp”/>
        <t id=”5” word=”pousos” lemma=”pouso” pos=”n”/>
        <t id=”6” word=”,” lemma=”--” pos=”pu”/>
        <t id=”7” word=”a” lemma=”a” pos=”prp”/>
        <t id=”8” word=”as” lemma=”o” pos=”art”/>
        <t id=”9” word=”8h50” lemma=”8h50” pos=”n”/>
        <t id=”10” word=”.” lemma=”--” pos=”pu”/>
    </terminals>
</s>
```

Source: Own elaboration.

[10] Computational tool that can analyze and identify the sentence constituents and their syntactic functions (CARROL, 2004).

[11] This is the case of *8:50,* which was wrongly annotated as noun.

[12] Each sentence (*s*) is annotated with two attributes: *id*, i.e., position of the sentence in the text, and *text*, i.e., the sentence itself of the *corpus*. The constitutive elements of *s* (words or expressions and punctuation symbols) are called *terminals* (or *tokens*). Each of them is described by four attributes: *id* (i.e., position in sentence), *word* (i.e., occurrence of the word or expression), *lemma* (i.e., canonical form) and *pos* (i.e., part-of-speech or word category).

[13] POS stands for *Part-of-Speech*.

We selected the subtopics of the previous annotation of CSTNews described in Cardoso et al. (2011). Tables 7 and 8 show the source texts of the sentences that illustrate the Historical background relation in Table 2. Based on the annotation[14], we may see, for example, that S1 from Table 2 is the S1 in Document 2 (Table 7) and its content conveys the subtopic labelled as "*the crash*", while S2 from Table 2 is the S4 in Document 1 (Table 8) and its content coveys a different subtopic, labeled as "*history*".

Table 7 – Example of subtopic annotation in CSTNews (Document 2, *Cluster* 1).

S1: A plane crash in Bukavu, in the Eastern Democratic Republic of the Congo (DRC) killed 17 people on Thursday, said a United Nations spokesman on Friday *(Um acidente aéreo na localidade de Bukavu, no leste da República Democrática do Congo (RDC), matou 17 pessoas na quinta-feira, informou nesta sexta-feira um porta-voz das Nações Unidas)* S2: The victims of the accident were 14 passengers and three crew members. *(As vítimas do acidente foram 14 passageiros e três membros da tripulação.)* S3: Everyone died when the plane, hampered by the bad weather, failed to reach the runway and crashed in a forest that was 15 kilometers from the airport in Bukavu. *(Todos morreram quando o avião, prejudicado pelo mau tempo, não conseguiu chegar à pista de aterrissagem e caiu numa floresta a 15 quilômetros do aeroporto de Bukavu.)* S4: The plane exploded and caught fire, said the UN spokesman in Kinshasa, Jean-Tobias Okala. *(Segundo fontes aeroportuárias, os membros da tripulação eram de nacionalidade russa.)* S5: The plane exploded and caught fire, added UN spokesman in Kinshasa, Jean-Tobias Okala. *(O avião explodiu e se incendiou, acrescentou o porta-voz da ONU em Kinshasa, Jean-Tobias Okala.)* S6: "There were no survivors", said Okala. *"Não houve sobreviventes", disse Okala.)* **<t LABEL="the crash" TOP= "1")>**
S7: The spokesman said the plane, a Soviet Antonov-28, of Ukrainian manufacturing and under ownership of the Trasept Congo, a Congolese company, also took a mineral load *(O porta-voz informou que o avião, um Soviet Antonov-28 de fabricação ucraniana e propriedade de uma companhia congolesa, a Trasept Congo, também levava uma carga de minerais.)* **<t LABEL= "detail about the aircraft" TOP= "3">**

Source: Own elaboration.

[14] The annotation of a subtopic includes two attributes: LABEL, i.e., a brief description of the subtopic, and TOP, i.e., a sequential number of annotation in the cluster (CARDOSO et al., 2012).

Table 8 – Example of subtopic annotation in CSTNews (Document 1, *Cluster* 1).

S1: At least 17 people died after a passenger plane crashed in Democratic Republic of Congo. *(Ao menos 17 pessoas morreram após a queda de um avião de passageiros na República Democrática do Congo.)* S2: According to a spokesman from UN, the plane was trying to land at the airport in Bukavu in the middle of a storm. *(Segundo uma porta-voz da ONU, o avião, de fabricação russa, estava tentando aterrissar no aeroporto de Bukavu em meio a uma tempestade.)* S3: The aircraft hit a mountain and crashed into a forest about 15 Km from the runway end. *(A aeronave se chocou com uma montanha e caiu, em chamas, sobre uma floresta a 15 quilômetros de distância da pista do aeroporto.)* **<t LABEL= “the crash” TOP= “1”>**
S4: Air accidents are frequent in Congo, where 51 private companies operate elderly planes built in the former Soviet Union. *(Acidentes aéreos são freqüentes no Congo, onde 51 companhias privadas operam com aviões antigos principalmente fabricados na antiga União Soviética.)* **<t LABEL= “historical background” TOP= “2”>**
S5: The crashed airplane, operated by Air Traset airline, carried 14 passengers and three crew members *(O avião acidentado, operado pela Air Traset, levava 14 passageiros e três tripulantes.)* S6: The airplane was a flight from the mining town Lugushwa, 130 kilometers away from its destination, Bukavu, *(Ele havia saído da cidade mineira de Lugushwa em direção a Bukavu, numa distância de 130 quilômetros.)* **<t LABEL= “the crashed aircraft” TOP= “1”>**
S7: Aircraft are used extensively for transport in Democratic Republic of Congo, a huge country where there are few paved road. *(Aviões são usados extensivamente para transporte na República Democrática do Congo, um vasto país no qual há poucas estradas pavimentadas.)* S8: In March, the European Union banned all Congolese airlines from operating in Europe. *(Em março, a União Européia proibiu quase todas as companhias aéreas do Congo de operar na Europa.)* S9: Only one kept the permission. *(Apenas uma manteve a permissão.)* S10: In June, the International Air Transport Association also included Congo in a group of several African countries it classed as an “embarrassment” to the industry. *(Em junho, a Associação Internacional de Transporte Aéreo incluiu o Congo num grupo de vários países africanos que classificou como “uma vergonha” para o setor.)* **<t LABEL= “historical background” TOP= “2”>**

Source: Own elaboration.

We selected the temporal expressions (TE) from the *corpus* annotation described in Menezes-Filho and Pardo (2011). In order to identify and classify the TEs of the source sentences from CSTNews, the authors used the typology of Baptista, Hagège and Mamede (2008). Such typology organizes TEs into 4 classes or types: (i) *calendar time* (i.e., the TE corresponds to a unique anchoring of the process onto the timeline), (ii) *duration* (i.e., the TE does not anchor the process onto the timeline) (e.g., *A reunião durará 2 horas*) (*The meeting will last 2 hours*), (iii) *frequency* (i.e., the TE relates the process to the timeline through multiple anchoring instances) (e.g., *Ocorrerá entre os dias 29 e 31 de julho*) (*It will take place between 29 and 31 July*), and (iv) *generic* (i.e., the expression does not anchor any event onto the timeline) (e.g., *Eu gosto do mês de julho*) (*I like July*). The calendar time class is further structured in: (i) *hour* (e.g., *Ele chegou às 9h30m*) (*He arrived at 9:30am*), (ii) *date* (e.g., *Em 7 de julho foi registrado 52.77 graus Celsius*) (O*n July 7th, 52.77 degrees Celsius were recorded*), and (iii) *interval* (i.e., TE involving two explicit dates (e.g., *A reunião durará entre 1 e 2 horas*) (*The meeting will last from 1 to 2 hours*). Finally, the *date* type is also classified based on the temporal reference of the TE and/or its indeterminacy regarding its anchoring on the timeline. In this sense, the following subtypes are distinguished: (i) *enunciation* (e.g., *Ele partiu em março*) (*He departed in March*)), (ii) *textual element*, somewhere in the text (e.g., *Um acidente no dia anterior*) (*A car accident the day before*), and (iii) *absolute dates*, directly computable from the TE (e.g., *Ele foi lançado em maio de 2009*) (*It was released in May 2009*). Based on such typology, the expression às 6h (*at 6am*) in (8) (see case (i) in Table 2) was annotated as *calendar time* of the subtype *hour*.

(8) A pista auxiliar de Congonhas abriu **<TE TYPE=”CALENDAR_TIME” SUBTYPE=”HOUR”>**às 6h **</TE>**, apenas para decolagens

We organized the linguistic information necessary to calculate the 7 attributes in an single *xls* file, which is here illustrated in two Tables (9 and 10) for matter of space. In Table 9, we exemplify the characteristics of the sentences showed in Table 2 that are needed for calculating the numeric attributes, i.e., NoI and Distance.

Table 9 – Pre-processing of the *corpus* to calculate the numeric attributes.

Corpus		Linguistic description	
Pair	CST Relation	Noun	Sentence position
1	Historical background	*acidente, localidade, Bukavu, leste, República Democrática do Congo, RDC, pessoa, porta-voz, Nações Unidas, quinta-feira, sexta-feira* (accident, city, Bukavu, eastern, Democratic Republic of the Congo, RDC, people, spokesman, United Nations, Thursday, Friday)	1
		acidente, Congo, companhia, avião, União Soviética (accident, Congo, company, plane, Soviet Union)	4
2	Follow-up	*pista, Congonhas, decolagem* (runway, Congonhas, departure)	6
		Congonhas, pouso (Congonhas, landing)	4

Source: Own elaboration.

In Table 10, we exemplify the description of the sentences that is needed for calculating the binary attributes (i.e., SubTol, TES1, TES2, ADVS1, and ADVS2). "X" indicates that the linguistic feature does not occur in the sentence.

Table 10 – Pre-processing of the *corpus* to calculate the binary attributes.

Corpus		**Linguistic description**		
Pair	CST Relation	Subtopic	Temporal Expression	Adverb
1	Historical background	1	calendar/date/enunciation	X
		2	X	X
2	Follow-up	1	calendar/hour	X
		1	calendar/hour	X

Source: Own elaboration.

After describing each sentence of the *subcorpus*, we manually calculated the attributes for each sentence pair. The values resulting from the attributes specification of the 90 sentence pairs in our *subcorpus* were also organized in an *xls* file, as shown in Table 11.

Table 11 – Two sentence pairs and their correspondent attributes and values.

Corpus		Attributes						
Pair	CST Relation	NoI	Distance	SubTol	TES1	TES2	ADVS1	ADVS2
1	Historical background	0.133	3	no	yes	no	no	no
2	Follow-up	0.4	2	yes	yes	yes	no	no

Source: Own elaboration.

In Table 11, the pair 1 is characterized by the attribute-value pair NoI=0.125, which means that the sentences have little information content in common. The value 0.125 for NoI results from the fact that among the 16 different nouns contained in S1 (i.e., *acidente, localidade, Bukavu, leste, República Democrática do Congo, RDC, pessoa, porta-voz, Nações Unidas, quinta-feira, sexta-feira*) and the 5 nouns contained in S (i.e., *acidente, Congo, companhia, avião, União Soviética*), only *acidente* occurs in both. Concerning Distance, we normalized[15] the values due to the different sizes of the source texts. Thus, the initial value 3 obtained for pair 1 in Table 11 was divided by the highest value of Distance observed in the *subcorpus* (i.e. 34), resulting in 0.088, the normalized value. Moreover, the sentence pair 1 do not share the same subtopic, which is indicated by the value "no" for the SubTol attribute. We also observe in Table 11 that there is no occurrence of adverbs in the pair 1, but temporal expressions, on the other hand, occur in S1 of the pair 1 (*na quinta-feira* and *nesta sexta-feira*).

Once the attributes for each sentence pair of the *subcorpus* were computed, we evaluated the potential of the attributes for automatic distinguishing Historical background from Follow-up.

Evaluation of the attributes

The attributes were automatically evaluated based on ML algorithms available in Weka (*Waikato Environment for Knowledge Analysis*) (HALL et al., 2009), an open source software toolkit from the University of Waikato (New Zealand). The toolkit supports both supervised and unsupervised ML algorithms from various Artificial Intelligence approaches.

In this paper, we used the supervised inductive learning approach. In such ML task, an algorithm learns from a training set (or *corpus*) whose classes (i.e., label to be learned) of the examples are known. In general, each example of a training data is a pair consisting of an input object (i.e., a sentence pair and its attributes) and a desired

[15] The normalization process aims at reducing the chances of data becoming inconsistent. Thus, a normalized attribute takes values in the range [0, 1].

output value or class (i.e., the CST relation of the pair). A supervised learning algorithm analyzes the training set and produces a *classifier* that should be able to predict the correct classes of new and unlabeled examples.

Specifically, we applied the *10-fold cross validation*[16] technique, which gets more realistic estimates of the error rates for classification, since our dataset is relatively small. To evaluate the attributes, we applied *precision, recall* and *f-measure*, i.e., metrics most commonly used to quantify the performance of NLP techniques (MITCHELL, 1997). *Precision* is the proportion of the instances (i.e., sentence pairs) returned by the classifier that were correct (i.e., instances classified as belonging to the correct class). *Recall* is the proportion of all possible correct instances that were returned. In other words, recall indicates the fraction of pairs annotated by human experts as belonging to a class *x* that were also correctly identified by the classifier. Finally, *f-measure* is the harmonic mean of the previous two measures. In this paper, we only used general accuracy for evaluating the results.

Although there are different ML paradigms, i.e., connectionist, mathematical (or probabilistic), and symbolic, we focused on symbolic algorithms, since they produce rules that can be easily interpreted and verified by human experts. Nonetheless, we have also tested other ML techniques from other paradigms, for comparison purposes only.

To identify the CST relations that codify complementarity, we tested the well-known connectionist method called *Multi-Layer Perception* (MLP), with the default Weka configurations. We achieved 82.2% of general accuracy. Among the several mathematical or probabilistic methods in Weka, we run Naïve-Bayes and SMO. Naive-bayes achieved the same accuracy of MLP (i.e., 82.2%), while SMO obtained 80% of accuracy.

Among the symbolic algorithms, we tried the same set used in previously related works, such as those of Maziero (2012), and Souza, Di-Felippo and Pardo (2012, 2013). Thus, we specifically tried One-R (or *One Rule*) (HOLTE, 1993), PART (WITTEN; FRANK, 1998), JRip (COHEN, 1995) e J48 (QUINLAN, 1993). One-R is probably the simplest symbolic algorithm, since it uses only the most discriminative feature or attribute to produce a single set of rules over this feature. JRip and PART are basic algorithms. They examine the classes from training data in growing size and generate an initial set of logical rules for the class. Such rules commonly combine two or more attributes. J48 builds decision trees from a set of training data. The most common way to build a decision tree is by top-down partitioning, and thus the most discriminative (or more generic) attribute corresponds to the topmost decision node and all other descendant nodes are less discriminative attributes.

J48 and PART built the biggest set of rules (13 and 8, respectively) with similar accuracy: both obtained 81% of general accuracy. One-R, as we mentioned, produces a single rule based on single attribute. In this case, the One-R algorithm selected TES2 as

[16] In k-fold cross-validation, the *corpus* is randomly partitioned into k equal sized subsamples. Of the k subsamples, a single one is retained for test, and the remaining (k 1) subsamples are used as training data. The process is repeated k times, with each of the k subsamples used once as the test data. The results are averaged over all the runs.

the most important attribute and, as usual, it surprisingly produced a very good result, obtaining 80% of accuracy. JRip, in its turn, generated the smallest set of rules (only 5) with general accuracy of 79%.

Thus, JRip learned a small set of rules with a very good level of accuracy. Such combination (i.e., manageable rule set and high accuracy among the symbolic approaches) makes the choice of JRip perfectly adequate for our purposes. Table 12 presents the 5 rules of JRip, which are followed by the number of instances (pairs) correctly classified and incorrectly classified, and the precision of the rule, given by the number of correctly classified instances over all the instances classified by that rule.

Table 12 – JRip logic rules for identifying temporal complementarity.

Rules	**Correct**	**Incorrect**	**Precision**
1. If TES2= hour, then *Follow-up*	12	0	100%
2. Elseif TES2= no, then *Follow-up*	30	8	78.9%
3. Elseif NoI>=0.315 e TES1=date-absolute, then *Follow-up*	4	0	100%
4. Elseif TES2=hour_date-enunciation, then *Follow-up*	3	0	100%
5. Elseif *Historical background*	41	4	91.1

Source: Own elaboration.

Based on these rules, one can say that the TES2, NoI and TES1 features define the pairs annotated as Follow-up in CSTNews. Among them, TES2 is the most discriminative attribute, since three of the total five rules are based on it. For *feature selection*[17], we applied the InfoGainAttributeEval algorithm, also available at Weka. This algorithm also indicated the relevance of TES2. In addition, one can say that, if none of the 4 first rules are applied, the default class or (i.e., CST relation) is Historical Background*,* which is given by rule 5.

It is also interesting to notice how productive the rules are. For instance, rules 1 and 2 deal with many more cases than rules 3 and 4, which is natural to happen due to the way the ML process chooses the attributes to start the rules. Thus, we can say that the classifier might still achieve good results by using only the two first rules (1 and 2) for Follow-up and the last default rule (5) for Historical background. In Table 13, we have the JRip confusion matrix, through which we verify in more details how the classifier is dealing with each class or relation. Each column of the matrix represents the instances (sentence pairs) in a predicted class (relation), while each row represents the instances in an actual class.

[17] Feature selection is the process of selecting a subset of relevant features. The aim of this process is to improve the performance of the classifiers. The central premise when using a feature selection technique is that the data contains features that are either redundant or irrelevant, and, removing them can reduce the processing time and generate simpler models.

Table 13 – Confusion matrix of JRip algorithm.

Class / **Test**	*Follow-up* (45)	*Historical background* (45)
Follow-up	35	10
Historical background	9	36

Source: Own elaboration.

It can be observed from the results in Table 13 that, from the total of 45 Follow-up pairs, the rules of JRip correctly classified 35 of them, and, from the total of 45 Historical background pairs, the algorithm correctly identified 36 of them. Based on this performance, we conclude that JRip correctly classified the pairs from both classes in a very similar way.

It is important to say that such results are only indicative of the features that characterize temporal complementarity and the corresponding CST relations, as well as the discriminative power of the attributes. We say that because our *subcorpus*, from which the classifiers were learned and tested, is very small. Apart from that, it is important to say that this is a pioneer study on the characterization of temporal complementarity as a linguistic phenomenon in MDS.

Future studies may include creating a testing *corpus*, by selecting new sentence pairs annotated as Follow-up and Historical background from CSTNews. Consequently, the *subcorpus* of 90 pairs analyzed here could be used as training data only, and the resulting classifier could be evaluated in a different set of sentence pairs. We expect that future efforts will be put in linguistic description of non-temporal complementarity and the corresponding CST relation (Elaboration).

Acknowledgments

We thank the Coordination for the Improvement of Higher Education Personnel (CAPES), and the State of São Paulo Research Foundation (FAPESP) for the financial support. We also thank the coordinator of the Interinstitutional Center for Computational Linguistics (NILC) for allowing us to use the laboratory facilities and for sharing considerable knowledge.

SOUZA, J.; Di FELIPPO, A. Caracterização da complementaridade temporal: subsídios para sumarização automática multidocumento. **Alfa**, São Paulo, v.62, n.1, p.121-147, 2018.

▪ *RESUMO: A complementaridade é um fenômeno multidocumento comumente observado entre notícias que versam sobre um mesmo evento. A partir de um corpus em português composto* *por um conjunto de pares de sentenças manualmente anotadas com as relações da Cross-Document Structure Theory (CST) que explicitam a complementaridade temporal (Historical background e Follow-up), identificou-se um conjunto potencial de atributos linguísticos desse tipo de complementaridade. Por meio de algoritmos de Aprendizado de Máquina, testou-se o potencial dos atributos em distinguir as referidas relações. O classificador simbólico gerado pelo algoritmo JRip obteve o melhor desempenho ao se considerar a precisão e o tamanho reduzido do conjunto de regras. Somente com base em 5 regras, tal classificador identificou Follow-up e Historical background com precisão aproximada de 80%. Ademais, as regras do classificador indicam que o atributo ocorrência de expressão temporal na sentença 2 é o mais relevante para a tarefa. Como contribuição, salienta-se que o classificador aqui gerado pode ser utilizado nos analisadores discursivos multidocumento para o português do Brasil que são baseados na CST.*

▪ *PALAVRAS-CHAVE: Descrição linguística. Complementaridade. CST. Sumarização Multidocumento. Processamento Automático de Língua Natural.*

REFERENCES

ALEIXO, P.; PARDO, T. A. S. CSTTool: um parser multidocumento automático para o Português do Brasil. In: IV Workshop on MSc Dissertation and PhD Thesis in Artificial Intelligence – WTDIA. 2008. **Proceedings**… Salvador/Brasil.

BAPTISTA, J.; HAGÈGE, C.; MAMEDE, N. Proposta de anotação e normalização de expressões temporais da categoria TEMPO para o HAREM II. In: Encontros do Segundo HAREM, 2008. **Actes**... p.1-24.

BICK, E. **The parsing system "PALAVRAS":** Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. 2000. 412 f. Thesis (PhD) - Aarhus University, Denmark University Press, 2000.

CARDOSO, P. C. F. **Exploração de métodos de sumarização automática multidocumento com base em conhecimento semântico-discursivo**. 2014. 182 f. Tese (Doutorado em Ciências da Computação) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2014.

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. CSTNews – A discourse-annotated *corpus* for single and multi-document summarization of news texts in brazilian portuguese. In: 3rd RST Brazilian Meeting, 2011. **Proceedings**… Cuiabá/Brazil. p.88-105.

CARROL, J. Parsing. In: MITKOV, R. (Ed.). **The Oxford handbook of computational linguistics.** Oxford/USA. Ed. Oxford University Express, 2004. p.233-248.

COHEN, W. Fast effective rule induction. In: 12th International Conference on International Conference on Machine Learning, 1995. **Proceedings**... California/USA. p.115-123.

CONDORAVDI, C.; CROUCH, D.; DE PAIVA, V.; STOLLE, R.; BOBROW, D. G. Entailment, intensionality and text understanding. In: HLT-NAACL 2003 workshop on Text meaning, 9., 2003. **Proceedings**... Edmonton/Canada: Association for Computational Linguistics, 2003. p.38-45.

ERKAN, G.; RADEV, D. R. LexPageRank: prestige in multi-document text summarization. In: Empirical Methods in Natural Language, 2004. **Proceedings**… Barcelona/Spain. p.365-371.

FELLBAUM, C. **WordNet:** an electronic lexical database. California: Ed. MIT Press, 1998.

HALL, M.; FRANK, E., HOLMES, G.; PFAHRINGER, B.; REUTEMANN, P.; WITTEN, I. H. The WEKA Data Mining Software: An Update. **SIGKDD Explorations**, v.11, Issue 1, 2009.

HATZIVASSILOGLOU, J. L.; KLAVANS, J. L.; HOLCOMBE, M.; BARZILAY, R.; MCKEOWN, K. Simfinder: a flexible clustering tool for summarization. In: NAACL Workshop on Automatic Summarization, 2001. **Proccedings**…Pittsburgh/USA. p.1-9.

HENDRICKX, I.; DAELEMANS, W.; MARSI, E.; KRAHMER, E. Reducing redundancy in multi-document summarization using lexical semantic similarity. In: 2009 Workshop on Language Generation and Summarisation, 2009. **Proceedings**... Suntec/Singapore: Association for Computational Linguistics, 2009. p.63-66.

HOLTE, R. C. Very simple classification rules perform well on most commonly used datasets. **Machine Learing**, Ed. Springer, v.11, n.1, p.63-90, 1993.

JORGE, M. L. C.; PARDO, T. A. S. Experiments with CST-based Multidocument Summarization. In: ACL Workshop TextGraphs-5: Graph-based Methods for Natural Language Processing, 2010. **Proceedings**… Uppsala/Sweden. p.74-82.

JURAFSKY, D.; MARTIN, J. H. **Speech and language processing:** an introduction to natural language processing, computational linguistics, and speech recognition. v.2. Englewood: Ed. Prentice Hall, 2009.

KOCH, I. G. V. **Introdução à linguística textual:** trajetória e grandes temas. 2. ed. São Paulo: Contexto, 2009.

KUMAR, Y. J.; SALIM, N.; RAZA, B. Cross-document structural relationship identification using supervised machine learning. **Applied Soft Computing**, v.12, n.10, p.3124-3131, oct. 2012.

LAGE, N. **Estrutura da notícia**. São Paulo: Ed. Ática, 2002.

MARNEFFE, M-C DE. **What's that supposed to mean? Modeling the pragmatic meaning of utterances**. 2012. 178 f. Thesis (PhD in Linguistics) – Department of Linguistics, Stanford University, Stanford, 2012.

MARNEFFE, M-C DE.; RAFFERTY, A. N.; MANNING, C. D. Finding contradictions in text. In: Annual meeting of the ACL, 46., 2008. **Proceedings**... Columbus/USA, p.1039-1047.

MANI, I. **Automatic Summarization.** Amsterdam/Netherlands: Ed. John Benjamins Publishing Company, 2001.

MANN, W. C.; THOMPSON, S. A. **Rhetorical structure theory**: a theory of text organization. California/USA: Ed. University of Southern California – Information Sciences Institute, 1987. p.87-190.

MAZIERO, E. G. **Identificação automática de relações multidocumento**. 2012. 117 f. Dissertação (Mestrado em Ciências da Computação) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2012.

MAZIERO, E. G.; JORGE, M. L. R. C.; PARDO, T. A. S. Revisiting Cross-document Structure Theory for multi-document discourse parsing. **Information Processing & Management**, v.50, n.2, p.297-314, 2014.

MENEZES FILHO, L. A.; PARDO, T. A. S. Detecção de Expressões Temporais no Contexto de Sumarização Automática. In: 2nd STIL Student Workshop on Information and Human Language Technology, 2011. **Proceedings**… Cuiabá/Brasil. p.1-3.

MITCHELL, T. M. **Machine learning**. v.45. Burr Ridge, IL: McGraw Hill, 1997.

MIYABE, Y.; TAKAMURA, H.; OKUMURA, M. Identifying cross-document relations between sentences. In: 3rd International Joint Conference on Natural Language, 2008. **Proceedings**… Hyderabad/India. p.141-148.

NEWMAN, E.; DOMN, W.; STOKES, N.; CARTHY, J.; DUNNION, J. Comparing redundancy removal techniques for multi-document summarization. In: Starting AI researchers' symposium, 2004. **Proceedings**…Valencia/Spain. p.223-228.

QUINLAN, J. **Programs for machine learning.** San Mateo/USA: Ed. Morgan Kaufmann Publishers, 1993.

RADEV, D. A common theory of information fusion from multiple text sources step one: cross-document structure. In: 1st SIGdial workshop on Discourse and dialogue, 10., 2000. **Proceedings**… Hong Kong/China: Association for Computational Linguistics, 2000. p.74-83.

RADEV, D. R.; JING, H.; BUDZIKOWSKA, M. Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation and user studies. In:

ANLP/NAACL Workshop on Automatic Summarization, 2000. **Proceedings...** Seatle/WA: North American Association for Computational Linguistics, 2000. p.21-29.

RADEV, D. R.; TEUFEL, S.; SAGGION, H.; LAM, W.; BLITZER, J.; QI, H.; CELEBI, A.; LIU, D.; DRABEK, E. Evaluation challenges in large-scale multi-document summarization: the mead project. In: 41st Annual Meeting of the Association for Computational, 2003. **Proceedings**… Sapporo/Japan: Association for Computational Linguistics, 2003. p.375-382.

RADEV, D.; OTTERBACHER, J.; ZHANG, Z. CSTNank: A corpus for the study of Cross-document Structural Relationship. In: **Proceedings fo Fourth International Conference on Language Resources and Evaluation**. Lisboa, 2004.

SOUZA, J. W. C. **Descrição linguística da complementaridade para a Sumarização Automática Multidocumento.** 2015. 105 f. Dissertação (Mestrado em Linguística) – Centro de Educação e Ciências Humanas, Universidade Federal de São Carlos, São Carlos, 2015.

SOUZA, J. W. C.; DI-FELIPPO, A.; PARDO, T. A. S. **Investigação de métodos de identificação de redundância para Sumarização Automática Multidocumento**. Série de Relatórios do NILC (NILC-TR-12): São Carlos/Brasil. 30 f. 2012.

________. Identificação da redundância na Sumarização Automática Multidocumento: explorando métodos superficiais. In: 3rd Student Workshop on Information and Human Language Technology (TILic), 2013. **Proceedings**… Fortaleza/Brasil. p.1-3.

SPARCK JONES, K. What might be in a summary? **Information Retrieval**. v.93, p.9-26, 1993.

TAUFER, P. Massa de informações digitais pode ser usada em benefício da população. **Jornal da Globo**, 26 dez. 2013. Disponível em: <http://g1.globo.com/jornal-da-globo/noticia/2013/12/massa-de-informacoes-digitais-pode-ser-usada-em-beneficio-da-populacao.html>. Acesso em: 02 fev. 2015.

WITTEN, I. H.; FRANK, E. **Generating accurate rule sets without global optimization.** Working paper series, 1998.

ZAHRI, N. A. H. B.; FUKUMOTO, F. Multi-document summarization using link analysis based on rhetorical relations between sentences. In: 12th International Conference on Computational Linguistics and Intelligent Text Processing, 2., 2011. **Proceedings**… Tokyo/Japan. p.328-338.

ZHANG, Z.; RADEV, D. Combining labeled and unlabeled data for learning cross-document structural relationships. In: **Natural Language Processing** – I JCNLP. Springer, p.32-41, 2005.

ZHANG, Z.; OTTERBACHER, J.; REDEV, D. R. Learning cross-document structural relationships using boosting. In: International Conference on Information and Knowledge Management, 2003. **Proceedings**… Las Vegas/USA. p.124-130.

ZHANG, Z.; BLAIR-GOLDENSOHN, S.; RADEV, D. R. Towards CST-enhanced summarization. In: Innovative applications of artificial intelligence conference, 2002. **Proceedings**… Edmonton/Canada. p.439-446.

Received in December 30, 2016

Approved in July 30, 2017