*Rev. Estud. Ling.*, Belo Horizonte, v. 29, n. 2, p. 859-907, 2021

# An investigation of linguistic problems in automatic multi-document summaries

## *Uma investigação de problemas linguísticos em sumários automáticos multidocumento*

Márcio de Souza Dias
Universidade Federal de Goiás (UFG), Catalão, Goiás / Brasil
marciosouzadias@ufg.br
http://orcid.org/0000-0003-1116-6965

Ariani Di Felippo
Universidade Federal de São Carlos (UFSCar), São Carlos, São Paulo / Brasil
arianidf@gmail.com
http://orcid.org/0000-0002-4566-9352

Amanda Pontes Rassi
Redação Nota 1000 Ltda., São Paulo, São Paulo / Brasil
amandarassi85@gmail.com
http://orcid.org/0000-0001-5314-1868

Paula Christina Figueira Cardoso
Universidade Federal de Lavras (UFLA), Lavras, Minas Gerais / Brasil
paula.cardoso@ufla.br
http://orcid.org/0000-0003-3621-8960

Fernando Antônio Asevedo Nóbrega
Samsung, São Paulo, São Paulo / Brasil
fernandoasevedo@gmail.com
http://orcid.org/0000-0002-1129-0133

Thiago Alexandre Salgueiro Pardo
Universidade de São Paulo (USP), São Carlos, São Paulo / Brasil
taspardo@icmc.usp.br
http://orcid.org/0000-0003-2111-1319

eISSN: 2237-2083
DOI: 10.17851/2237-2083.29.2.859-907

**Abstract:** Automatic summaries commonly present diverse linguistic problems that affect textual quality and thus their understanding by users. Few studies have tried to characterize such problems and their relation with the performance of the summarization systems. In this paper, we investigated the problems in multi-document extracts (i.e., summaries produced by concatenating several sentences taken exactly as they appear in the source texts) generated by systems for Brazilian Portuguese that have different approaches (i.e., superficial and deep) and performances (i.e., baseline and state-of-the art methods). For that, we first reviewed the main characterization studies, resulting in a typology of linguistic problems more suitable for multi-document summarization. Then, we manually annotated a corpus of automatic multi-document extracts in Portuguese based on the typology, which showed that some of linguistic problems are significantly more recurrent than others. Thus, this corpus annotation may support research on linguistic problems detection and correction for summary improvement, allowing the production of automatic summaries that are not only informative (i.e., they convey the content of the source material), but also linguistically well structured.

**Keywords:** automatic summarization; multi-document summary; linguistic problem; corpus annotation.

**Resumo:** Sumários automáticos geralmente apresentam vários problemas linguísticos que afetam a sua qualidade textual e, consequentemente, sua compreensão pelos usuários. Alguns trabalhos caracterizam tais problemas e os relacionam ao desempenho dos sistemas de sumarização. Neste artigo, investigaram-se os problemas em extratos (isto é, sumários produzidos pela concatenação de sentenças extraídas na íntegra dos textos-fonte) multidocumento em Português do Brasil gerados por sistemas que apresentam diferentes abordagens (isto é, superficial e profunda) e desempenho (isto é, métodos *baseline* e do estado-da-arte). Para tanto, as principais caracterizações dos problemas linguísticos em sumários automáticos foram investigadas, resultando em uma tipologia mais adequada à sumarização multidocumento. Em seguida, anotou-se manualmente um corpus de extratos com base na tipologia, evidenciando que alguns tipos de problemas são significativamente mais recorrentes que outros. Assim, essa anotação gera subsídios para as tarefas automáticas de detecção e correção de problemas linguísticos com vistas à produção de sumários automáticos não só mais informativos (isto é, que cobrem o conteúdo do material de origem), como também linguisticamente bem-estruturados.

**Palavras-chave:** sumarização automática; sumário multidocumento; problema linguístico; anotação de *corpus*.

Submitted on April 29th, 2020

Accepted on July 1st, 2020

## 1 Introduction

Multi-document Summarization (MDS) is an important area of Natural Language Processing (NLP). It aims at automatically producing a unique summary for a set of source texts on the same topic (MANI, 2001; NENKOVA; MCKEOWN, 2011). It currently has attracted a lot of attention in the scientific community because of the increasing incredible amount of available textual information nowadays, mainly on the web.

It is a consensus that a good summary should contain the most relevant information in the texts, and the area has achieved significant progress in producing summaries that are more informative. The progress is the result of both linguistically poor and rich summarization methods, such as the empirical/statistical approaches (see, e.g., ANDO *et al*., 2000; CARBONELL *et al*., 1997; HAGHIGHI; VANDERWENDE, 2009; MIHALCEA; TARAU, 2005; RIBALDO *et al*., 2016) and the deep ones (CARDOSO; PARDO, 2016; CASTRO JORGE; PARDO, 2010; MCKEOWN; RADEV, 1995; RADEV, 2000; ZHANG *et al*., 2002).

Automatic summaries must also present the information to the reader in a cohesive and coherent way. According to Koch (1998), cohesion is related to the surface organization of a text. It may be expressed by successive links among elements in the superficial structure of the text. For example, anaphoric pronouns, which refer back to textual antecedents, are elements of cohesion. Coherence is related to the meaning of a text; related to the possible interpretation of the text (KOCH; TRAVAGLIA, 2002). Beaugrande and Dressler (1981) claim that the continuity of meaning is what keeps the text coherent. Thus, coherence is the combination of concepts and relations of textual elements and, sometimes, it is necessary to make use of world knowledge and knowledge about the interlocutors and the situation itself for the text to make sense. For example, coherence can be created between sentences through repetition of words, which helps to reiterate the same ideas.

Although current summarization methods are still limited on such aspects, since most of the systems only produce extractive[1] instead of abstractive summaries[2] (which are still hard to achieve and not fully understood, systematized and formalized). Trying to evaluate

[1] Summaries produced by concatenating sentences taken exactly as they appear in the source texts.

[2] Summaries that allow rewriting operations over the original material.

the linguistic quality (LQ) of summaries through numeric scores using lexical, syntactic and/or semantic features (see, e.g., CONROY *et al.*, 2011; GIANNAKOPOULOS; KARKALETSIS, 2011; LIN *et al.*, 2012; OLIVEIRA, 2011; PITLER *et al.*, 2010) or to identify certain problematic linguistic aspects (see, e.g., CRISTINI; DI-FELIPPO, 2019; FONSECA *et al*., 2019; FRIEDRICH *et al.*, 2014; PITLER *et al.*, 2010), the summarization literature has revealed that automatic extracts present several problems that affect their LQ.

In order to propose specific solutions for improving the LQ of automatic summaries or more sophisticated MDS methods that tackle such issues it is necessary to identify and to characterize the problems in a corpus of automatic summaries.

In this paper, we investigate the types of LQ problems that affect multi-document summary quality. Initially, we reviewed the main approaches in the literature of linguistic problems in automatic summaries, resulting in a typology more suitable for the multi-document scenario. Next, we used the typology to annotate a corpus of extractive multi-document summaries in Brazilian Portuguese[3] produced by systems with different performances, from both superficial (that use little linguistic knowledge) and deep approaches (which are based on sophisticated linguistic knowledge, as semantics and discourse), including baseline and state-of-the art methods. Finally, with the annotated corpus, we systematized and characterized the problems that the systems produce and show that some problems are significantly more recurrent than others.

In Section 2, we present an overview of basic concepts in multi-document summarization, focusing on the methods used to produce the summaries that we evaluated. Section 3 presents the linguistic problems that are available in the literature, resulting in a typology of problems. In Section 4, we present our corpus of summaries used in the annotation. In Section 5, we detailed the annotation of linguistic problems in the multi-document summaries in Portuguese. Section 6 shows the results and the analysis of the error annotation. In Section 7, the final remarks will be presented.

[3] The LQ problems are generic and may be applied to any language.

## 2 Automatic Summarization

In this section, we present an overview of basic concepts in Automatic Summarization and methods developed specifically for generating summaries in Brazilian Portuguese.

### 2.1 Basic concepts

According to Mani (2001), a summary is a shorter version of one or more texts. Depending on the number of documents to summarize, the automatic process is defined as single or multi-document summarization. While the first dates back to the 50s, the latter, which is the focus of this paper, consists in a more recent initiative that officially started in the 90s, bringing new challenges to the Automatic Summarization area.

There are several possible classifications for summaries (see, e.g., MANI; MAYBURY, 1999). Summaries may be informative, indicative or critical. Informative summaries include the main facts of the source documents organized in a cohesive and coherent way. These summaries can be read in place of the original texts. Indicative summaries, differently from the informative ones, do not substitute the original texts, but only indicate what the texts are about. For example, indexes may be classified as indicative summaries. Critical summaries bring the authors' opinions or points of view about the source texts. Examples of critical summaries are book reviews.

Summaries are also classified according to the intended audience. Generic summarization does not take into account any specific interest of the reader, producing general-purpose summaries. On the other hand, summarization focused on the interest of the reader uses information based on his/her prior knowledge and interests. For example, a layman may need a summary with more contextual information about the subject, while a reader with a good knowledge about the subject may expect that the summary presents additional or new information.

Summaries may be classified as extractive or abstractive. Extractive summaries are formed by pieces of non-modified text, with copy and paste operations (from the source texts to the summaries), basically. Abstractive summaries make use of rewriting operations, i.e., there is some or full modification in the structure and/or in the writing of the source text passages for building the corresponding summaries. Currently, most of the available automatic summarizers are extractive since abstraction is still considered a very difficult task.

The construction of summaries may follow two linguistic approaches: superficial/shallow and deep approaches (MANI, 2001). Shallow approaches use little or no linguistic knowledge at all to produce summaries. The main advantage of the shallow approach is its robustness[4] and scalability,[5] but it may produce worse summaries than the ones resulting from deep approaches. Deep approaches use linguistic knowledge, theories and formal language models in the creation of summaries, as lexicons, wordnets, grammars, and syntactic-semantic and discourse analysis. This approach is considered the most complex one, because of the number of linguistic variables. Its application is usually limited since systems of this approach are mostly developed for specific domains. Shallow and deep approaches may also be merged, resulting in the hybrid approach.

Finally, another important concept in summarization is the amount of information that will be included in the summaries, which is determined by the compression rate, i.e., the ratio between the size of the summary and the size of the source texts (MANI, 2001), usually measured in number of words.

In this paper, we conduct our investigation with extractive, informative and generic summaries (which consist in the most usual configuration in the area), produced by both shallow and deep approaches for Portuguese. We briefly introduce the main characteristics of the summarization methods that we used in what follows.

### 2.2 Summarization methods for Portuguese

There are several multi-document summarization systems for Portuguese, following different content selection strategies, using both classical and state of the art methods in the area. For this investigation, we have selected four of them, trying to get a sample of summaries of different performances, which represent the main available approaches.

One of them was GistSumm (GIST SUMMarizer) (PARDO *et al*., 2003; PARDO, 2005). This summarizer follows a simple shallow approach, and, to the best of our knowledge, it was the first one made available for Portuguese. Its approach is based on the gist of the source

[4] In this case, a robust method is applicable to very different testing data, e.g., different genre or domain.

[5] The scalability represents the ability of the method to deal with large amount of data.

texts, i.e., the main idea intended to be conveyed or understood by the reader. The gist is the most important segment of the source texts, commonly expressed by only one sentence. The most widely applied technique for detecting it has been simple word frequency measures. Once identified, the gist serves as guide for identifying and selecting other sentences to compose the final extract. Figure 1 shows a summary generated by GistSumm.

FIGURE 1 – Summary generated by GistSumm

[S1] The crimes happened in the city of Muttur, in which during the last two weeks, there were severe conflicts between the troops of the Sri Lanka army and the guerrillas of the Liberation Tigers of Tamil Eelam (LTTE).

[S2] The director of ACF in Sri Lanka, Benoit Miribel, confirmed the death of its employees and said that the NGO "did not suffer a similar loss in over 25 years of existence."

[S3] The violent conflict started on July 26, when government air troops bombed positions of the guerrillas after the rebels blocked a dam located in its territory for more than a week, hindering the supply of water in places under the government control.

[S4] The special envoy for the peace in Sri Lanka from Norway, Jon Hanssen-Bauer, arrived in the island last week and met the two parties, attempting to reduce the tension and to avoid a new start of the civil war.

[S5] The crimes happened in the city of Muttur, in which, during the last two weeks, there were severe conflicts between the troops of the Sri Lanka army and the guerrillas of the Liberation Tigers of Tamil Eelam (LTTE).

[S6] The director of ACF in Sri Lanka, Benoit Miribel, confirmed the death of its employees and said that the NGO "did not suffer a similar loss in over 25 years of existence."

[S7] The special envoy for the peace in Sri Lanka from Norway, Jon Hanssen-Bauer, arrived in the island last week and met the two parties, attempting to reduce the tension and to avoid a new start of the civil war.

[S8] Fifteen local employees of a French charity institution in Sri Lanka were found dead in the city of Muttur in the north of the country.

One may see that the summary has several problems, such as redundant information (S1 with S5, S2 with S6, and S4 with S7), noun phrases without explanation (e.g., "the crimes" in S1 is not specified or explained), and acronyms without explanation ("ACF" and "NGO" in S2). Such problems occur due to the simplicity of GistSumm, whose

method is considered a baseline method for Portuguese. It was included in this investigation for historical reasons and to evidence improvements and remaining problems that the best current methods show.

The RSumm summarizer (RIBALDO *et al.*, 2012, 2016) is based on classical graph-based methods, which use the relationship map approaches of Salton *et al.* (1997) adapted for MDS. According to the authors, graphs/maps are built from a set of documents on the same topic, where each vertex represents a sentence and the edges indicate the lexical similarity between the sentences. The best method groups topic-related sentences and select the most relevant one from each subtopic to compose the summary. Figure 2 shows an example of a summary generated by RSumm for the same source texts of the summary in Figure 1. One may see that problems still happen in the summary, mainly related to the proper introduction of noun phrases. However, it is clear that this summary is much better than the one produced by GistSumm.

FIGURE 2 – Summary generated by RSumm

[S1] The special envoy for the peace in Sri Lanka from Norway, Jon Hanssen-Bauer, arrived in the island last week and met the two parties, attempting to reduce the tension and to avoid a new start of the civil war.

[S2] Fifteen local employees of a French charity institution in Sri Lanka were found dead in the city of Muttur in the north of the country.

[S3] The crimes happened in the city of Muttur, in which, during the last two weeks, there were severe conflicts between the troops of the Sri Lanka army and the guerrillas of the Liberation Tigers of Tamil Eelam (LTTE).

Cardoso and Pardo (2015, 2016) presented a deep method for MDS. They assume that the relevance of a sentence is influenced by its salience in its source text, which is given by Rhetorical Structure Theory (RST) (MANN; THOMPSON, 1987), using the method proposed by Marcu (1999), and its salience in the set of texts, given by Cross-document Structure Theory (CST) (RADEV, 2000). The method is referred by RC-4 (which stands for the "**4**$^{th}$ combination of **R**ST and **C**ST information"). Figure 3 shows a summary generated by RC-4.

FIGURE 3 – Summary generated by RC-4

[S1] Fifteen volunteers from the French NGO "Action Contre la Faim" (ACF) were killed in northeastern Sri Lanka today, said a spokeswoman

[S2] According to a representative of the group Action Contre la Faim, the bodies were found in the organization office.

[S3] The director of ACF in Sri Lanka, Benoit Miribel, confirmed the death of its employees and said that the NGO did not suffer a similar loss in over 25 years of existence.

[S4] Up to now, the Sri Lankan authorities did not confirm the deaths or clarified what happened in the city of Muttur.

[S5] The rebels said that they will consider a new bombing of the army.

This summary is much better than the others, but it still presents some problems, such as lack of connection between the S5 content and the rest of the summary, and occurrence of the noun phrases "The rebels" and "a new bombing of the army" that do not have their respective referents in the summary.

The last summarizer is based on a statistical method (CASTRO JORGE, 2015). It captures summarization patterns by estimating the occurrence probability of some features in human summaries, including, e.g., discourse (following the RST and CST models) and sentence position information. The features represent strategic characteristics that indicate the salience of a sentence among a set of sentences. The probabilistic model is based on a generative learning approach (the noisy-channel framework), where the task is formulated with probabilistic components, including probabilities for content selection during the transformation process and for coherence of the produced summary, and a decodification step (i.e., the production of the final summary). This summarization method is referenced by MTRST-MCAD (Method of Transformation with RST and Model for Coherence evaluation After Decodification). Figure 4 shows an example of a summary created by the MTRST-MCAD method.

FIGURE 4 – Summary generated by MTRST-MCAD

[S1] It is unclear who committed the murders of the employees of the French organization.

[S2] The rebels said that they will consider a new bombing of the army.

[S3] Up to now, the Sri Lankan authorities did not confirm the deaths or clarified what happened in the city of Muttur.

[S4] "We tried to send a team to Muttur to check what is going on, but the soldiers did not allow us to enter the city, which is totally blocked", he said.

[S5] The director of ACF in Sri Lanka, Benoit Miribel, confirmed the death of its employees and said that the NGO did not suffer a similar loss over 25 years of existence.

One may see that the summary also has some problems that affect its quality, such as the lack of connection between S2 content and the rest of the summary, and the occurrence of the definite noun phrases "the murders of the employees" and "the French organization" in S1 that do not have their respective referents. The same occurs with the definite noun phrase "The rebels" and "the army" in S2. Besides these problems, the explanations for the "ACF" and "NGO" acronyms in S5 are not present in the summary.

The RC-4 system (in the deep approach) is currently the best method for Portuguese, followed very closely by RSumm (in the shallow approach). With some distance, we have MTRST-MCAD and, finally, GistSumm. The evaluations of these methods have so far been guided by summary informativeness criteria, mainly using ROUGE (LIN, 2004), a standard n-gram-based measure that is automatically computed, allowing for fast and easily reproducible evaluation. Despite the importance of informativeness, the examples in this section show that this criterion is not enough for assuring that good summaries are produced and provide evidence that the systems need to treat problems that affect the LQ of their summaries, as they severely harm the summary quality. For this, we believe that the definition and the identification of problems related to LQ will guide the summarizers in possible solutions for these problems.

In what follows, we present and discuss important issues and previous initiatives related to defining and characterizing linguistic problems in summaries, proposing, in the end, a synthesized and comparative view of them. This forms the basis of the study that we conduct in our corpus.

## 3 Definition and characterization of linguistic problems

Some works have tried to find and deal with linguistic problems in summaries for improving their quality. Although some identified problems are similar, some approaches are much more refined than others and there is great variation in the error catalogues. To the best of our knowledge, we briefly list and discuss the main initiatives in what follows.

### 3.1 The revision of linguistic quality issues in automatic summaries

Otterbacher *et al.* (2002) studied the problems related to the cohesion of extractive multi-document summaries and suggested revisions (solutions) to improve cohesion. The authors presented a corpus-based analysis of automatically generated extractive multi-document summaries, produced by the MEAD summarizer (RADEV *et al.*, 2003), which is one of the most popular summarization systems for English. The authors discussed the feasibility of automatically improving the summaries and they created a taxonomy of problems related to cohesion.

According to them, the taxonomy is divided into five pragmatic categories related to textual cohesion in multi-document summaries: *Discourse*, *Identification of Entities*, *Temporal Expressions*, *Grammar*, and *Location Settings*. In what follows, we detail these problems and some of their main related problems, showing examples.

The *discourse* category focuses on the relationships among the sentences of the summary (inter-sentence level) and on the relationships among textual elements inside sentences (intra-sentence level). The authors considered some aspects in this category that may cause cohesion problems in multi-document summaries: *Topic Shift*, *Lack of Purpose*, *Contradiction*, *Redundancy*, and *Conditional Sentences*.

The *Topic Shift,* which is the fast change of one subject by another, has the highest occurrence (45%). In order to solve the problem, an addition of a transitional sentence or phrase may be necessary, as illustrated in Figure 5. The underlined segment is a possible example of transitional phrase in a *Topic Shift.*

FIGURE 5 – Example of solution for *topic shift* problem

| |
|---|
| [S1] <u>In a related story</u>, the government of Hong Kong announced a proposal to require all drug rehabilitation centers... |

Source: Otterbacher *et al*. (2002)

Another common problem in summaries is sentences with *lack of purpose*, which may be solved by the addition of sentences or phrases that motivate a purpose in the problematic segment. Figure 6 shows this situation.

FIGURE 6 – Example of solution for *lacked purpose*

| |
|---|
| [S1] In order to assist the ongoing investigation as the cause of the crash, the U.S. team from the National Transportation Safety Board will join experts... |

Source: Otterbacher *et al*. (2002)

*Contradiction* is related to some information in a given sentence that contrasts with one or more previous sentences. In such cases, a discourse marker such as "however" or "in contrast" may help. Figure 7 shows an example of contradiction.

FIGURE 7 – Example of *contradiction* that was solved

| |
|---|
| [S1] However, according to reports on CNN, the control tower was concerned with the speed and altitude of the plane and had discussed these concerns with the pilot. |

Source: Otterbacher *et al*. (2002)

*Redundancy* occurs when a sentence contains previously reported information. For Otterbacher *et al*. (2002), a possible action to solve this problem is to delete the redundant constituent (non-head element of NPs, PPs, or the entire relative clause or phrase). Figure 8 shows an example of this scenario, where the underlined passage must be removed.

FIGURE 8 – Example of *redundancy* that may be solved

| |
|---|
| [S1] The crash of flight 072 that killed 143 people… |
| [S2] The plane, which was carrying the 143 victims, was headed for Bahrain from Egypt. |

Source: Otterbacher *et al*. (2002)

According to the authors, sometimes events in a given sentence are *conditioned* by events in another sentence. Thus, a good action is to modify the sentences, using the structure "IF (sentence 1), (sentence 2)".

Besides this, the verb tenses may be changed to represent the condition. Figure 9 is an example of this use.

FIGURE 9 – Example of *conditional sentence* with improved cohesion

| |
|---|
| [S1] If the proposed measures were implemented, they would ensure broadly the same registration standard to be applied to all drug treatment centers. |

Source: Otterbacher *et al*. (2002)

The *identification of entities* category requires the resolution of referential expressions, since the reader needs to identify each entity mentioned in a summary. According to Otterbacher *et al*. (2002), 9 problems were found in summaries related to this category, which were: *Underspecified Entity*, *Misused Quantifier*, *Overspecified Entity*, *Repeated Entity*, *Bare Anaphora*, *Misused Definite Article*, *Misused Indefinite Article*, *Missing Article*, and *Missing Entity*. The *underspecified entity* problem was the most frequent in this category, in 38% of the cases.

The authors also use some revisions to solve problems related to the identification of entities. For example, one possible solution to solve an *underspecified entity* (a newly mentioned entity that has no description, or the presence of an acronym without explanation) is the addition of a full name, a description or a title for the new entity, or expanding the acronym if this is the case. Figure 10 shows an example of this revision.

FIGURE 10 – Example of *underspecified entity* revision

| |
|---|
| [S1] Mrs. Clarie Lo, the Commissioner of Narcotics, said the proposal would be introduced to non-medical drug treatment centers. |

Source: Otterbacher *et al*. (2002)

The *misused definite article* problem may also be solved by adding a definite article if the entity has already been mentioned, or an indefinite article if the entity is new. Figure 11 shows part of a text with the addition of the indefinite article "*a*", since the entity "*second eruption"* is new in the text.

FIGURE 11 – Example of *misused definite article* revision

| |
|---|
| [S1] On Thursday, a second eruption appeared to be smaller than anticipated. |

Source: Otterbacher *et al*. (2002)

The *temporal* category is related to the right temporal relationships among events. The authors identified five types of possible problems that fall into this category: *Temporal Ordering*, *Time of Event*, *Event Repetition, Synchrony* and *Anachronism*. The *temporal ordering* problem represented 89% of all errors found in this category.

*Temporal ordering* is related to the establishment of correct temporal relations among events. If there is a problem, the authors recommend, e.g., to add time expressions, to add ordinal numbers, to delete inappropriate time expressions, or to modify an existing time expression. Figure 12 shows an example of a temporal ordering problem that was revised.

FIGURE 12 – Example of revision for *temporal ordering* error

| |
|---|
| [S1] Two days later, a second eruption appeared to be smaller than scientists had anticipated. |

Source: Otterbacher *et al*. (2002)

The *event repetition* problem may be solved by simply adding an adverb such as "again". Figure 13 shows an example of such revision.

FIGURE 13 – Example of *event repetition* problem revision

| |
|---|
| [S1] Mount Pinatubo is likely to explode again in the next few days or weeks. |

Source: Otterbacher *et al*. (2002)

Some problems in *grammar* category have also been identified in the corpus used by Otterbacher *et al*. (2002) Among these problems are: *Run-on Sentence*, *Mismatched Verb*, *Missing Punctuation*, *Awkward Syntax*, *Parenthetical*, *Subheadings/Titles*, and *Misused Adverb*. The *run-on sentence* problem was the most frequent one, representing 35% of these errors.

For the authors, a *run-on sentence* is a very long sentence. Thus, the authors recommend splitting long sentences into two separate sentences and deleting the conjunction. Figure 14 shows a long sentence that was revised.

FIGURE 14 – Example of *run-on sentence* problem revision

| |
|---|
| [S1] Lt. Col. Ron Rand announced at 5 a.m. Monday that all personnel should begin evacuating the base.<br>[S1] Meanwhile, dawn skies over central Luzon were filled… |

Source: Otterbacher *et al*. (2002)

*Parenthetical* is a problem related to the inappropriate use of parenthesis. Thus, the authors simply suggest deleting the parenthesis symbols. Figure 15 shows an example of inappropriate use of parenthesis.

FIGURE 15 – Example of a *parenthetical* problem revised

| |
|---|
| [S1] (Volcanoes such as Pinatubo arise where one of the earth's crust plates is slowly diving beneath another.) |

Source: Otterbacher *et al*. (2002)

The *location settings* category includes a type of revision related to the correct location of events, in order for the text to be improved. These settings may be: *Location of Event*, *Collocation*, *Change of Location*, and *Place/Source Stamp*.

*Location of event* specifies where an event takes place. Thus, the authors suggest adding a prepositional phrase that indicates place (city, state, or country). Figure 16 shows a type of *location of event* setting that was revised.

FIGURE 16 – Example of a *location of event* setting revision

| |
|---|
| [S1] Three bodies were lain before the faithful in the Grand Mosque <u>in Manama, Bahrain</u> during a special prayer… |

Source: Otterbacher *et al*. (2002)

*Collocation* is related to two or more events that occur in the same place. Thus, the authors suggest adding a prepositional phrase or an adverb that indicates the collocation. An example is shown in Figure 17.

FIGURE 17 – Example of revision for *collocation*

| |
|---|
| [S1] Meanwhile, <u>in the same area</u>, search teams sifted through the wreckage. |

Source: Otterbacher *et al*. (2002)

Generally, according to the authors, the *discourse* category corresponded to 34% of all the problems found in the corpus, followed by the categories *identification of entities* (with 26%), *temporal expressions* (22%), *grammar* (12%), and *location settings* (6%).

Friedrich *et al*. (2014) presented a corpus of multi-document summaries (called LQVSumm) which was manually annotated with several types of LQ errors. These summaries were automatically created in the TAC (Text Analysis Conference) 2011 shared task on Guided Summarization (OWCZARZAK; DANG, 2011). The authors identified two classes of problems: one considering entity mentions and another happening at the level of clauses. The first is related to reference or coreference problems. The last involves grammar or redundancy errors.

For the authors, in the level of entity, the problem types are: *First mention without explanation*, *Subsequent mention with explanation*, *Definite noun phrase without reference to previous mention*, *Indefinite noun phrase with reference to previous mention*, *Pronoun with missing antecedent*, *Pronoun with misleading antecedent*, and *Acronyms without explanations*.

The *first mention without explanation* problem is assigned to the first mention of an entity for which there is not a clear reference to the reader. For example, in the sentence "Paul bought toys to the poor children", there is no sufficient introduction for the entity "Paul".

The *subsequent mention with explanation* problem is related to entity mentions that have already been referenced in the text and present an inappropriate extra explanation. For example, consider sentences S1 and S2 in Figure 18. In sentence S2, there is an additional explanation related to the entity Taylor, but the entity has already been referenced in sentence S1.

FIGURE 18 – Example of *subsequent mention with explanation* error

| |
|---|
| [S1] Taylor's attorney could not be reached for comment Friday night. |
| [S2] Tony Taylor, 34, of Hampton, Va., has a plea-agreement hearing scheduled for 9 a.m. |

Source: Friedrich *et al*. (2014)

The *definite noun phrase without reference to previous mention* problem occurs when a definite noun phrase is used to refer to the first

mention of an entity in the text. For example, "the Petrobras Company" should be used in a summary in which "a company" has been mentioned before.

The *indefinite noun phrase with reference to previous mention* error occurs when an indefinite noun phrase is used for an entity already mentioned in the discourse. For example, the noun phrase "*a company*" is not appropriate if the same company has already been mentioned in the summary.

The *pronoun with missing antecedent* problem occurs when there is no possible antecedent that matches with the pronoun. Figure 19, for example, shows a beginning of an automatic multi-document summary where the pronoun "he" does not have a possible antecedent.

FIGURE 19 – Example of *pronoun with missing antecedent*

| |
|---|
| [S1] The renouncement may not stop the investigation because the process was already started. |
| [S2] He will establish the process against the deputies involved with the Sanguessugas Mafia. |

Source: Cardoso *et al*. (2011)

The *pronoun with misleading antecedent* error occurs when an anaphoric expression refers to a misleading antecedent and its right antecedent is not in the summary. For example, Figure 20 shows part of a summary about soccer. In this case, the pronoun "he" (in the second sentence) apparently refers to the soccer player Kaká (in the first sentence), but, in the source text, the pronoun refers to Robinho, who is not introduced in the summary.

FIGURE 20 – Example of *pronoun with misleading antecedent*

| |
|---|
| [S1] At the 27 minutes, Kaká kicked the ball and Ronaldinho diverted the kick. |
| [S1] 20 cm from the end line, he gave two humiliating dribbles in the Ecuadorian defender and crossed the ball to Elano, who scored the fourth goal, at 37 minutes. |

Source: Cardoso *et al*. (2011)

The *acronyms without explanations* problem occurs when acronyms are not previously known and are not explained in the first time they are introduced.

Friedrich *et al*. also proposed the annotation at the clause level. This was made on arbitrary spans, from single tokens to complete sentences. According to the authors, the clause level errors are: *Incomplete sentence*, *Inclusion of datelines*, *Other ungrammatical form*, *No semantic relatedness*, *Redundant information*, and *No discourse relation*.

An *incomplete sentence* problem usually results from segmentation errors in sentence compression (or truncation), which aims at reducing the length of candidate sentences to generate summaries with the desirable size pre-defined by the compression rate. For example, the following sentence is incomplete, since the name of the person was lost in the end of the sentence: "One was killed in a bedroom and others were murdered in a classroom, according to the head of the campus police, W."

For the authors, the *inclusion of datelines* in summaries is not desired and should be avoided. For example, a summary with the information "GEORGETOWN, Pennsylvania 2006-10-05 16:53:53 UTC" must be annotated with this problem.

The *other ungrammatical form* error considers all other ungrammaticality cases, such as missing spaces and wrong punctuation.

The *no semantic relatedness* problem occurs when sentences do not show plausible semantic relations. In Figure 21, for example, S1 and S2 are apparently not related.

FIGURE 21 – Example of *no semantic relatedness* problem

| |
|---|
| [S1] It is popularly known as the 'pink city' because of the ochre-pink hue of its old buildings and crenellated city walls. |
| [S2] He said there was no justification for such killings. |

Source: Friedrich *et al*. (2014)

The *redundant information* problem occurs when two or more sentences express the same information. For example, in Figure 22, sentences S1 and S2 are partially redundant.

FIGURE 22 – Example of summary with *redundant information*

| |
|---|
| [S1] The suspect apparently called his wife from a cell phone shortly before the shooting began, saying he was "acting out in revenge for something that happened 20 years ago", Miller said. |
| [S2] The gunman, a local truck driver Charles Roberts, was apparently acting in "revenge for an incident that happened to him 20 years ago. |

Source: Friedrich *et al*. (2014)

The *no discourse relation* problem, in particular, may happen when an explicit discourse connective (e.g., "and", "but", "even though" and "because") is no longer appropriate in the new context in the summary, does not being suitable for signaling the corresponding discourse relation. For example, this is the case for the connective "and" in the second sentence in Figure 23.

FIGURE 23 – Example of *no discourse relation* error in a summary

| |
|---|
| [S1] Taylor's attorney could not be reached for comment Friday night. |
| [S2] And the person who cooperates first gets the biggest reward. |

Source: Friedrich *et al*. (2014)

It their conclusions, the authors show that there are relationships between the types of problems they defined and the summary readability evaluation performed at TAC, which we introduce in what follows.

In the mono-document summarization, Kaspersson *et al*. (2012) investigated linguistic problems that occur in summaries extracted from single texts. The focus was on discourse problems, such as referring expressions with missing antecedents and fragments, and how text units in the summaries are connected. In addition, the authors have investigated how the different size of summaries and different genres influence the occurrence of types of problems. The authors considered texts of three different genres in their study: Swedish newspapers, popular Swedish science texts, and authority texts from the Swedish Social Insurance Administration.

The problems found by the authors were grouped into three categories: *Erroneous anaphoric reference*, *Absent cohesion or context*, and *Broken anaphoric reference*. *Erroneous anaphoric reference* is related to an anaphoric expression in the summarized text that refers to an erroneous antecedent, given that the correct antecedent was not extracted from the source text of the summary. This category occurs for the following cases: *Noun phrases*, *Proper names*, and *Pronouns*. *Absent cohesion or context* is a self-explanatory error, related to the lack of cohesion or necessary context in summaries. *Broken anaphoric reference* happens when an anaphoric expression presented in a summary does not have its antecedent because this antecedent was not extracted

from the source text. This category also occurs for the following cases: *Noun phrases*, *Proper names*, and *Pronouns*.

The authors report that the most significant problems are: *Erroneous anaphoric reference related to pronoun*, *Absent cohesion or context*, *Broken anaphoric references related to noun phrases* and *Broken anaphoric references related to pronouns*.

For evaluating summaries in summarization contests, TAC (DANG, 2005) developed classical guidelines to evaluate LQ in summaries related to 5 features: *Grammaticality*, *No Redundancy*, *Referential Clarity*, *Textual Focus*, and *Textual Structure and Coherence*.

*Grammaticality* verifies whether there are format and grammar problems in the summaries, including capitalization (e.g., whether proper names start with a capital letter). In relation to *no redundancy*, a good summary should present the maximum amount of unique information that is possible in respect to the compression rate. Thus, a summary is weighted by the unnecessary repetition of information. This analysis must happen in different levels, such as the redundant data/fact of an event, sentences, and names (entities should be, whenever possible, referenced by pronouns). A summary presents *referential clarity* when text references are not ambiguous. A summary has *focus* when all sentences are related to the addressed issue. The last feature of TAC suggests that a summary is also evaluated by its good *structuring and coherence*. For example, a summary should not present divergent information on the same fact or event.

These 5 criteria that were proposed in TAC (actually, when it was named Document Understanding Conference (DUC)) are widespread in the area and used by most of the works that attempt to check LQ in summaries.

### 3.2 A synthesized view of linguistic quality issues

In section 4.1, we reviewed the more important sets of LQ problems in automatic summaries defined by previous research. Such sets present similarities and differences in several aspects, such as (i) coverage, since some problem sets are more complete than others; (ii) types of problems, (iii) generality of the problems (since some problem sets are more fine-grained than others), and (iv) purpose (some errors are tailored for single summarization, others are for MDS, and others are

more agnostic). This shows the relevance and the complexity of these studies, which support summarization and other tasks.

In Table 1, we synthesized the LQ problem sets, showing the similarities and differences based on 5 classes: (i) errors related to inappropriate formatting and metadata inclusion; (ii) problems with grammatical origin; (iii) inadequacies that come from style/grammar choices; (iv) problems related to inadequacies in the use of entities and, therefore, also related to cohesion; and (v) errors related to discourse and coherence. We indicate with an "X" when a study treats the respective LQ issue.

It is clear that some problem types cause problems in other levels (e.g., a grammar error of missing subject/agent in a sentence also results in lower cohesion), but we focused on the origin of the problems when categorizing them. It is also interesting to notice that such categorization may not be completely fair to the listed works, as they report different problem specificity levels: while Otterbacher *et al*. (2002) and Friedrich *et al*. (2014) present much more refined error catalogues, Kaspersson *et al*. (2012) and Dang (2005) are more worried with general level problems.

TABLE 1– Synthesis of LQ problems in summaries

| **LQ problems** | **Otterbacher *et al*. (2002)** | **Kaspersson *et al*. (2012)** | **Friedrich *et al*. (2014)** | **Dang (2005)** |
|---|---|---|---|---|
| **Formatting, metadata** | | | | |
| *inclusion of subheading/titles* | x | | | |
| *inclusion of place/source stamp* | x | | | |
| *inclusion of datelines* | | | x | x |
| *inclusion of system-internal formatting* | | | | x |
| **Grammar** | | | | |
| *missing subject/agent* | x | | x | x |
| *mismatched verb* | x | | x | x |
| *missing punctuation* | x | | x | x |
| *wrong parenthetical* | x | | x | x |
| *incomplete sentence* | | | x | x |
| *wrong capitalization* | | | | x |

| | | | | |
|---|---|---|---|---|
| **Grammar, style** | | | | |
| *run-on sentence* | x | | | |
| *awkward syntax* | x | | | |
| *missing/omitted article* | x | | | |
| **Entities, cohesion** | | | | |
| *first mention without explanation* | x | x | x | x |
| *acronyms without explanations* | x | x | x | x |
| *subsequent mention with explanation* | x | | x | x |
| *repeated entity* | | | x | x |
| *definite noun phrase without reference to previous mention* | x | x | x | x |
| *indefinite noun phrase with reference to previous mention* | x | x | x | x |
| *misused quantifier* | x | x | | x |
| *pronoun with missing antecedent* | x | x | x | x |
| *noun phrase with missing antecedent* | | x | | x |
| *proper noun with missing antecedent* | | x | | x |
| *pronoun with misleading antecedent* | | x | x | x |
| *noun phrase with misleading antecedent* | | x | | x |
| *proper noun with misleading antecedent* | | x | | x |
| *not clear identification of who or what the pronouns and noun phrases are referring to* | | | | x |
| **Discourse, coherence** | | | | |
| *occurrence of redundancy* | x | | x | x |
| *occurrence of contradiction* | x | | | x |
| *not explicit conditional sentences* | x | x | | x |
| *lack of purpose for a sentence* | x | x | | x |
| *lack of place specification for an event (including collocation, change of location)* | x | x | | x |

| | | | | |
|---|---|---|---|---|
| *lack of time specification for an event (including anachronism, temporal ordering, synchrony, repetition of event)* | x | x | | x |
| *abrupt topic shift* | x | x | | x |
| *no semantic relatedness* | x | x | x | x |
| *misused word/discourse marker* | x | x | x | x |

For multi-document processing tasks (as MDS), the last two problem types ("Entities, cohesion" and "Discourse, coherence") look more worthy of identification and treatment, as they are more frequent errors and cause more serious problems. Thus, as described in section 5 (specifically in section 5.1), we have based our corpus annotation on these LQ problems, looking for a more appropriate and informative error set for MDS.

In next section, we introduce the summarization corpus that we used to conduct our investigation of linguistic problems, over which we ran the above summarization methods and performed the corpus analysis.

## 4 The Corpus

The corpus used in this work was the CSTNews corpus (CARDOSO *et al.*, 2011). This corpus has been specially created for multi-document summarization. It is composed of 140 texts (with an average of 334 words and 14.9 sentences per text) distributed in 50 sets/clusters of news texts written in Brazilian Portuguese[6] from various domains. Each cluster has 2 or 3 texts from different sources that address the same topic. These sources are important Brazilian online newspapers, as *Folha de São Paulo*, *Estadão*, *O Globo*, *Jornal do Brasil*, and *Gazeta do Povo*.

According to the authors, the choice of these news agencies was due to their popularity, to publish the main current news, to the use of a clear and everyday language, and because they make available different

[6] The adoption of a corpus in Portuguese was due to the facts that (i) it was possible to have access to the several different summaries that we needed for this investigation and (ii) the annotators were native speakers of this language, which allowed for a more refined and reliable annotation.

versions of the same facts, which is important for a multi-document corpus.

Besides the original texts, the corpus contains several linguistic annotation layers, manually produced by experts, with satisfactory annotation agreement results. The manual annotations include single and multi-document summaries, text-summary alignments, the identification of temporal expressions, RST and CST annotation, noun and verb senses, segmentation of the source texts in subtopics, and semantic annotation of informative aspects in summaries, among other annotations. There are also some automatic annotations, which include morphosyntactic and syntactic analyses, with the best parser for Portuguese, and multi-document summaries.

For the annotation task, 200 multi-document summaries have been used since each of the four automatic summarizers generated one extract for each cluster of the CSTNews. Table 2 shows the average of words and sentences per summary generated by each summarizer.

TABLE 2 – Basic counts for the corpus of automatic summaries

| **System** | **Average of words** | **Average of sentences** |
|---|---|---|
| GistSumm | 362 | 11 |
| RSumm | 134 | 4 |
| RC-4 | 132 | 4 |
| MTRST-MCAD | 139.78 | 7.92 |

According to the table, the average of words and sentences in the summaries from GistSumm is higher than the summaries produced by the other summarizers. This happens because GistSumm compression rate is computed in a different way in relation to the other summarizers. It is computed over all the source texts, which are concatenated. For the other summarizers, the compression rate is 30% of the largest text of each cluster of the CSTNews corpus. We kept GistSumm in the comparison because we considered it interesting to see how the summary size variance affects the occurrence of LQ problems.

## 5 Annotation of linguistic problems in multi-document summaries

In this section, we describe the methodology that we used for the annotation of LQ problems in our corpus of automatic multi-document summaries in Portuguese. Such annotation allowed us to understand and categorize the linguistic problems, to check the quality of the automatic summaries and to guide the future development of automatic methods that judge the LQ of multi-document summaries and, consequently, of automatic summarizers.

Given the 50 clusters of the CSTNews corpus, the automatic summarizers GistSumm (PARDO *et al*., 2003; PARDO, 2005), RSumm (RIBALDO *et al*., 2012, 2016), RC-4 (CARDOSO; PARDO, 2015, 2016) and MTRST-MCAD (CASTRO JORGE, 2015) were used to generate individual extractive summaries for each cluster. Therefore, 200 automatic multi-document summaries were produced and manually annotated.

Based on the related literature and the analysis in section 4.2, we synthetically list the linguistic problems of interest in three categories: (i) *Entity Level*, (ii) *Clause Level*, and (iii) *Others* (see TABLE 3). In general, the problems we adopted are strongly based on those of Friedrich *et al*. (2014), extended with some more information and problem types that were necessary for our corpus annotation.

All errors were identified in the corpus with XML markers. The markers have the format <e TYPE=*(error name)*>*(Text Passage)*</e>. For some markers, there is additional information placed after the error name, and this will be explained along with their respective errors. The "*error name*" field is filled with the name of the error identified in the "*text passage*" field, which may contain full sentences or sentence fragments that show the error.

In what follows, the errors are explained once more, now adapted to this work and accompanied by the markup strategy and actual examples of our corpus.

### 5.1 The LQ problems typology

For the investigation of the problems in automatic multi-document extracts in Portuguese, we organized the linguistic problems of interest in 3 categories: (i) *Entity Level*, (ii) *Clause Level*, and (iii)

*Others* (errors that are different from the two first categories) (TABLE 3). The *Entity* and *Clause* categories have several types.

TABLE 3 – The typology of LQ problems

| Level | Problem type | Tag |
|---|---|---|
| ***Entity*** | *First mention without explanation* | 1M-EXP |
| | *Subsequent mentions with explanation* | SM+EXP |
| | *Definite noun phrase without reference to the previous mentions* | DNP-REF |
| | *Indefinite noun phrase with reference to the previous mentions* | INP+REF |
| | *Pronouns without antecedent* | PRO-ANT |
| | *Pronouns with misleading antecedents* | PRO_MIS |
| | *Acronyms without explanation* | ACR-EXP |
| ***Clause*** | *Redundant information* | RED |
| | *Contradiction* | CONTR |
| | *Incomplete sentences* | INC_SENT |
| | *No semantic relationship* | No_SEM |
| | *Connective/discursive marker without appropriate context* | DM |
| ***Other*** | *Errors that are different from the two first categories* | OTHER |

### 5.1.1 Problems in the entity level

Based on Table 3, one sees that the errors in the *entity level* present 7 subcategories: 1M-EXP, SM+EXP, DNP-REF, INP+REF, PRO-ANT, PRO_MIS, and ACR-EXP.

*First mention without explanation* (1M-EXP) is identified in a summary when the first mention of an entity is not properly introduced. In Figure 24, there is a problem of 1M-EXP in the third sentence (S3) of the summary. In this case, the first mention of entity "Tepco" was annotated because the reader does not know what this entity is, i.e., there is not a clear introduction to this entity in its first mention.

FIGURE 24 – Annotation of a1M-EXP problem

[S3]<e TYPE=1M-EXP>Tepco</e> has declared the earthquake did not cause leaks, but, afterwards, it revealed that 1,200 liters of water with radioactive material from the factory have leaked to the sea.

*Subsequent mentions with explanation* (SM+EXP) are identified in summaries when entities have already been mentioned in the text, but they still appear with an inappropriate (usually, extra) explanation. For illustration, consider sentences S1 and S2 in Figure 25.

FIGURE 25 – Annotation of a SM+EXP problem

[S1] The president of the Ethics Council of the Senate, Leomar Quintanilha (PMDB-TO), said to be contrary to the unification of the processes against the Senator Renan Calheiros (PMDB-AL).

[S2] <e TYPE=SM+EXP SENT=S1 TEXT= "The president of the Ethics Council of Senate, Leomar Quintanilha (PMDB-TO)"> The president of the Ethics Council of Senate, Leomar Quintanilha (PMDB-TO)</e>, said that he is against the union of representations, however that he will propose to a vote.

The entity "Leomar Quintanilha (PMDB-TO)" is explained in sentence S1 as "The president of the Ethics Council of Senate", and sentence S2 contains the same entity with a repeated explanation, characterizing a type SM+EXP problem. This problem is annotated in the second occurrence of the entity with explanation, as shown in Figure 25. The SENT field contains the identification of the sentence in which the first mention of the entity that was specified in the field TEXT occurs.

*Definite noun phrase without reference to previous mentions* (DNP-REF) is identified in summaries when a definite noun phrase does not refer to any entity mentioned earlier. For example, consider sentences S1, S2 and S3 in Figure 26.

FIGURE 26 – Example of a DNP-REF problem

[S1] At least 17 people died after the crash of a passenger plane in the Democratic Republic of Congo.

[S2] According to an ONU spokeswoman, the plane, Russian-made, was trying to land in the Bukavu airport in the midst of a storm.

[S3] <e TYPE=DNP-REF>The spokesman</e> informed that the plane, a Soviet Antonov-28 of Ukrainian-made and owned by a Congolese company, Trasept Congo, also carried a cargo of minerals.

The error <e TYPE=DNP-REF> in sentence 3 is due to the definite noun phrase "The spokesman", for which there is no reference to any entity mentioned earlier.

The *indefinite noun phrase with reference to previous mentions* (INP+REF) problem is identified in summaries when an indefinite article is used together with an entity already mentioned in the discourse (that, therefore, should be introduced in another way). For example, S2, in Figure 27, includes the indefinite noun phrase "an Airbus A320", which was already introduced in S1 ("The Airbus-A320"), causing inconsistency in the summary.

FIGURE 27 – Example of the INP+REF problem

[S1] In São Paulo, on Tuesday (17), the Airbus-A320 of TAM presented a defect in the reverse of the right turbine for the last 13 days.

[S2] The problem would have been detected by the electronic system of the plane, but the plane, <e TYPE=INP+REF SENT=S1 TEXT= "the Airbus-A320"> an Airbus A320</e>, continued flying with the right reverse off.

*Pronoun without antecedent* (PRO-ANT) is identified when a pronoun does not have a possible antecedent in the summary. For example, the first sentence of the summary in Figure 28 contains the pronoun "he" without a possible antecedent for it.

FIGURE 28 – Example of the PRO-ANT problem

[S1] Hospitalized in a hospital in Buenos Aires, <e TYPE = PRO-ANT>he</e> relapsed and started to feel pain again due to acute hepatitis, according to his personal doctor, Alfredo Cahe.

[S2] "Maradona had a relapse in acute hepatitis. Now, he is stable. Although he improved on Sunday, it is expected that he continues in hospital," Cahe declared to "La Nación".

*Pronoun with misleading antecedent* (PRO_MIS) is identified when an anaphoric expression refers to a misleading antecedent and its correct antecedent is not present in the summary. In this annotation task, the annotators could check the source text to identify the correct antecedent. In the example in Figure 29, the pronoun "he" (in S2) seems to connect to the entity "Kaká" (in S1). However, in the source text, the pronoun refers to the soccer player "Robinho", who is not cited in the summary.

FIGURE 29 – Example of the PRO_MIS problem

[S1] At 27 minutes, Kaká kicked from far away and Ronaldinho diverted the kick.
[S2] 20 cm from the end line <e TYPE=PRO_MIS ANT="Kaká, Ronaldinho">he</e> dribbled the Ecuadorian defender and crossed the ball to Elano, who scored the fourth goal at 37 minutes.

Besides identifying the type of error in the TYPE tag, the misleading antecedents must also be listed in the ANT tag. This allows the recovery of the problems in future studies.

*Acronyms without explanation* (ACR-EXP) are identified in a summary by their "non expanded form" or when they are not explained. For example, in the sentences in Figure 30, the "Deic" and "PF" acronyms have no proper introduction.

FIGURE 30 – Example of the ACR-EXP problem

[S1] The other suspect is graffiti man and, according to <e TYPE=ACR-EXP>Deic</e>, he has been arrested for theft, but has already been released.
[S2] The <e TYPE = ACR-EXP CS = "Federal Police"> PF </ e> did not know how to inform if this kind of reward is paid to law enforcement agencies.

Some acronyms are considered to be common sense, such as abbreviations of states and national (Brazilian) political parties. Such cases was annotated with the CS tag, which contains the common sense meaning of the acronym, as shown in the annotation of the error in Figure 30. In this work, common sense was used when the majority of the annotators had the same knowledge about the acronym. Differently from us, Friedrich *et al*. (2014) considered as common sense entities that are in a pre-compiled list of well-known acronyms.

### 5.1.2 Problems in the clause level

Based on Table 3, the *clause* category has 5 types of problems, which are: RED, CONTR, INC_SENT, No_SEM, and DM.

*Redundant information* (RED) (in total or partial levels) negatively affects the informativity of summaries. As an example, it is possible to see that sentence S2 in Figure 31 contains information from sentence S1, i.e., it is a repetition. Due to this, we marked this problem as a RED error in the TYPE tag, and we indicated the first sentence where the original information was present.

FIGURE 31 – Example of the RED problem

[S1] A homemade bomb was thrown against the building of the Public Ministry, in the center of the capital, but nobody was injured.

[S2] <e TYPE=RED SENT=S1> A homemade bomb exploded outside the building of the State Public Ministry and nearby shops were hit by shrapnels. </e>

*Contradiction* (CONTR) is identified when there is a conflict of information between two sentences. In Figure 32, sentences S1 and S2 have contradictory information in relation to the number of injured and dead people. Thus, we marked the sentence that presented the contradiction as CONTR, and we identified the sentence that presented the contradiction in the SENT tag.

FIGURE 32 – Example of the CONTR problem

[S1] The Egyptian Minister of Health Hatem, El-Gabaly, said on Monday that 57 people died and 128 were injured in the collision between two passenger trains in the Nile Delta, north of Cairo.

[S2] <e TYPE=CONTR SENT=S1> At least 80 people died and over 165 were injured on Monday after the collision of two passenger trains in the Nile Delta, north of Cairo, according to the police and the medical sources. </e>

*Incomplete sentence* (INC_SENT) is identified when there are no punctuation marks, space or complement of a sentence. For example, in the summary in Figure 33, sentence S2 finished with a comma, i.e., this sentence is considered incomplete.

FIGURE 33 – Example of the SENT_INC problem

[S1] As expected, the athlete Fabiana Murer won the gold medal in the pole vault at the Pan American Games in Rio, on Monday, at the João Havelange Stadium.

[S2] <e TYPE=INC_SENT>Murer won the highest place of the podium with the 4m60 mark against 4M40 of the American April Steiner.</e>

*No semantic relationship* (No_SEM) is identified when adjacent sentences do not present proper semantic relationship. As an example, Figure 34 contains a summary, in which there is not a clear relation between S2 and S1.

FIGURE 34 – Example of the No_SEM problem

[S1] Abadia was arrested in a residence located in a luxury condominium of Aldeia da Serra, in São Paulo.
[S2] <e TYPE=No_SEM>Four safes were also sealed</e> [...]

*Connective/discursive marker without appropriate context* (DM) is identified when the use of explicit discourse markers (e.g., "but", "because", "however") are considered inappropriate in the context of the summary. In the summary in Figure 35, the discourse marker "But" does not relate to the previous sentence. This happens due to the extractive nature of the summaries, which may include sentences without their contexts of occurrence. In the annotation of this error, we used the CONEC tag to identify the marker that is inappropriately used.

FIGURE 35 – Example of the DM problem

[…] [S4] Until the end of the game, Bruno and Anderson did not enter the court anymore.
[S5] <e TYPE=DM CONEC="But">But, after that, everybody in the gymnasium screamed the lifter name.</e> [...]

### 5.1.3 Other problems

In case of problems that were not listed in the previous categories, we labeled them as *Other* and the "EXPLANATION" tag contains the explanation of the error. For example, Figure 36 presents a summary that is problematic because it uses terms in different languages referring to the same entity/event (the "championship"), i.e., "Brasil Open" in sentence S1 and "Aberto do Brasil 2013" in sentence S2.

FIGURE 36 – Example of the *Other* problem: inappropriate references

[S1] In addition to Rafael Nadal, the tournament will have three more athletes among the 20 best of ATP ranking: the Spanish Nicolás Almago (11th place and 3 times champion of Brasil Open), the Argentinian Juan Mônaco (12th) and the Swiss Stanilas Wawrinka (17th).
[S2] The organization of <e TYPE=Other EXPLANATION="reference in Portuguese for the term introduced in English">Aberto do Brasil 2013</e> announced this Tuesday morning that the Spanish Rafael Nadal will be returning to the tournament to be disputed in February.

Problems as “Metadata inclusion” and “Distinct spelling for the same entity” are also considered as belonging to *Other*. Figures 37 and 38 show the respective examples for these problems.

FIGURE 37 – Example of the *Other* problem: metadata inclusion

<e TYPE=Other EXPLANATION=”Metadata inclusion”>FIRST -</e>Murer jumps to break the Pan American record; first gold medal in Athletics.

FIGURE 38 – Example of the *Other* problem: distinct spelling for the same entity

[S1] Israeli military forces in south of Lebanon also reported that, on Sunday, 30 militants of Hesbollah were killed, while an officer and two soldiers were wounded in Oiled.

[S2] The Israeli air force attacked 150 targets early this morning in Lebanon as the Jewish state soldiers killed 10 <e TYPE=Other EXPLANATION=”Distinct spelling for the same entity”>Hezbollah</e> militiamen in the Bint Djebeil and Kafr Hula Lebanese villages, according to military sources.

### 5.2. The task of linguistic problem annotation

The goal of the annotation was to identify the linguistic errors of the typology described in section 5.1 (see Table 3) in summaries that were automatically generated by the 4 cited automatic summarizers.

The task was carried out by a group of experts in a face-to-face process, i.e., it happened every day at a specific time and place for 1 hour. We believe that: 1 hour a day made the task less exhausting for the annotators and this may have positively influenced the annotation quality; everyday annotation, in turn, creates commitment to the task. The task was also better managed with all annotators in the same place.

We used some days to train the 6 annotators (2 linguists and 4 computational scientists) and to refine the guidelines with them. These annotators have been chosen because of their experience in NLP and with annotation tasks.

Due to the subjectivity of the task, the linguistic problems were only marked after a consensus among the annotators or when the majority of them agreed. This strategy is interesting because it produces a more consistent and correct annotation, allowing a more robust annotation with high linguistic error coverage. On the other hand, the annotation time is longer in comparison to the traditional strategies, in which each annotator

works with different summaries per day. In this work, the duration of the annotation task was approximately 150 days.

Some problems are interesting to comment. The *No semantic relationship* error was the error that required more attention and refinement in its interpretation, due to the high degree of subjectivity involved in this problem identification. Thus, this interpretation involved discussions among annotators until the reconciliation process, i.e., the final decision for marking the problem, as suggested by Hovy and Lavid (2010). The *Acronym without explanation* problem required that every annotator had the same background knowledge in order to fill the CS (common sense) field. This background knowledge may be different among the annotators and this may cause the inadequate identification of the problem. Therefore, the annotation approach used in this work may have avoided this type of problem.

Even with all the annotators working together, we periodically verified the agreement among them. In such case, each annotator separately worked with the same summaries, and, after this, we calculated the agreement by the Kappa measure (CARLETTA, 1996). Kappa is a classic agreement measure in NLP, which indicates the correlation between annotators while it discounts the agreement by chance. In the literature, there are some suggestions that guide the decision on the minimum agreement value that is expected: a value less than 0.4 may indicate an unreliable annotation; if it is between 0.4 and 0.75, the annotation is satisfactory; and if it is higher than 0.75, it is very good. This value, however, changes according to the subjectivity of the phenomenon and the difficulty of the annotation task. We consider our annotation task as a very difficult and subjective one. Thus, we expect lower kappa values.

We present the results of the annotation in the following section.

## 6 Results and discussion

### 6.1 Performance of the summarizers

For the 4 multi-document summarizers considered in this task (GistSumm, RSumm, RC-4, and MTRST-MCAD), 1,359 linguistic problems were identified. Table 4 shows the quantity of errors by summarizer.

TABLE 4 – Total of problems annotated for each summarizer

| **Systems** | **Annotated problems** | |
|---|---|---|
| | **Quantity** | **%** |
| GistSumm | 521 | 38.33 |
| MTRST-MCAD | 421 | 30.97 |
| RC-4 | 220 | 16.20 |
| RSumm | 197 | 14.50 |

As expected, Table 4 shows that there are more problems in the summaries produced by GistSumm than in the summaries of the others, which looks natural, given that GistSumm is a very simple summarizer and produces longer summaries than the other systems, running more risk to commit problems.

The statistics computed from our annotation show that *redundant information* (RED) is the most recurrent error, with a total of 261 occurrences in the summaries of the four summarizers (see TABLE 5). This result confirms that detection and properly treatment of redundancy are problematic issues in MDS. Together with *acronyms without explanation* (ACR-EXP) and *definite noun phrase without reference to the previous mentions* (DNP-REF), RED accounted for more than 50% of the problems.

TABLE 5 – Problems by subcategory

| **Problem subcategory** | **Qty.** | **%** |
|---|---|---|
| *Redundant information* (RED) | 261 | 19.20 |
| *Acronyms without explanation* (ACR-EXP) | 255 | 18.76 |
| *Definite noun phrase without reference to the previous mentions* (DNP-REF) | 182 | 13.39 |
| *Subsequent mentions with explanation (*SM+EXP) | 152 | 11.18 |
| *No semantic relationship* (No_SEM) | 136 | 10.00 |
| *Other* | 123 | 9.05 |

| | | |
|---|---|---|
| *First mention without explanation* (1M-EXP) | 103 | 7.57 |
| *Contradiction* (CONTR) | 41 | 3.01 |
| *Connective/discursive marker without appropriate context* (DM) | 37 | 2.72 |
| *Pronouns without antecedent* (PRO-ANT) | 30 | 2.20 |
| *Indefinite noun phrase with reference to the previous mentions* (INP+REF) | 25 | 1.83 |
| *Incomplete sentence* (INC_SENT) | 11 | 0.80 |
| *Pronouns with misleading antecedents* (PRO_MIS) | 3 | 0.29 |

Table 6 illustrates the quantity of *redundant information* (RED) error for each summarizer, as it is the most recurrent problem.

TABLE 6 – Total of *redundant information* (RED) problems

| **Systems** | **Problems** | |
|---|---|---|
| | **Quantity** | **%** |
| GistSumm | 160 | 61.30 |
| RC-4 | 55 | 21.08 |
| MTRST-MCAD | 23 | 8.81 |
| RSumm | 23 | 8.81 |

Redundancy errors may also directly increase the problems of the *entity* category as redundancy may cause repetitions and introduction of entities in an inappropriate way. For example, Figure 39 shows part of a summary with *redundant information* (RED) and problems related to the *entity* category embedded in the redundant sentences.

The sentences with repeated information (as in S5, S7 and S17) present errors of the *entity* category. In this case, for each redundant information error, there is one INP+REF error. This also certainly contributes to the high amount of annotated errors in the summaries produced by GistSumm.

FIGURE 39 – A GistSumm summary with *redundancy* caused by an *entity* category error

[S1] A new series of criminal attacks was recorded early on Monday, the 7th, in São Paulo and municipalities in the countryside of the State of São Paulo.

[S2] A homemade bomb was thrown against the building of the Public Ministry, in the state capital.

[S3] The criminal actions may have been ordered by the leaders of the Primeiro Comando da Capital (PCC), which had promised to return the attacks in São Paulo on Father's Day on Sunday.

[S4] At ABC Paulista, at least ten buses were set on fire - seven in Mauá and three in Santo André.

[S5] <e TYPE=RED SENT=S2><e TYPE=INP+REF SENT=S2 TEXT="A homemade bomb"> A homemade bomb </e> was thrown against <e TYPE= SM+EXP SENT=S2 TEXT = "the Public Ministry"> the Public Ministry (MP)</ e> headquarters. </e>

[S6] The building of the Treasury secretary, in the center, was hit by three homemade bombs.

[S7] <e TYPE=RED SENT=S3> The leaders of the criminal gang PCC had promised <e TYPE=INP+REF SENT=S1 TEXT="A new series of criminal attacks"> A new wave of attacks </e> will happen if the Public Ministry of São Paulo deny the temporary exit of prisoners because of Father's Day. </e> [...]

[S17] <e TYPE=RED SENT=S1,S3> Members of PCC had promised <e TYPE=INP+REF SENT=S1 TEXT="A new series of criminal attacks"> a new wave of attacks </e> will happen if the Public Ministry of São Paulo deny the temporary exit of prisoners because of Father's Day.</e> [...]

In relation to the quantity of problems by category, Table 7 synthesizes the achieved results. The *entity* category included the most frequent problems, which occurred 750 times. The fact that this category had the highest amount of problems was expected, since there are more entities than sentences in a summary. As an example, the summary in Figure 40 was generated by RSumm, and it does not present errors of the *clause* category. However, five annotated errors are related to the *entity* category, and 1 to the *other* category.

According to Table 7, the RC-4 and RSumm summarizers, which make use of more linguistic knowledge, present a lower quantity of errors than the others. In particular, the RSumm summarizer had the lowest quantity of annotated errors in two of the three categories; in the remaining category, it was outperformed by the RC-4 system only.

TABLE 7 – Quantity of problems by category

| **Systems** | ***Entity Level*** | ***Clause Level*** | ***Other*** |
|---|---|---|---|
| GistSumm | 239 | *221* | 61 |
| MTRST-MCAD | 252 | 129 | 40 |
| RC-4 | 123 | 83 | 14 |
| RSumm | 136 | 53 | 8 |
| **Total** | **750** | **486** | **123** |

FIGURE 40 – Example of Rsumm summary with more problems of the *entity* category

[S1] <e TYPE=DNP-REF>In the second round</e>, the vote intentions for President Lula fell from 53% in June to 50% in July, while candidate Alckmin increased from 29% to 36%.

[S2] <e TYPE = ACR-EXP>CNI</e> explains that the research does not provide a comparison with the previous survey for the first round, because it is the first time that <e TYPE=ACR-EXP> Ibope </e> uses the official list of candidates for president.

[S3] Although it does not allow comparisons, it is worth remembering that, in June, Lula had 48% of the votes; Alckmin 18% and <e TYPE=1M-EXP>Heloisa Helena </e> 5%.

[S4] The margin of error is two percentage points upwards or downwards.

[S5] <e TYPE=Other EXPLANATION="Phrase with ambiguous referent">The research</e> was held between 29 and 31 July and was registered in <e TYPE=ACR-EXP>TSE</e> under number 12197/2006.

Some important data is also presented in Table 8, such as the percentage of the problems that were found in the summaries generated by each of the four summarizers. We show in bold some of the main errors for each system.

According to the table, *redundant information* is the main problem in 2 of the 4 summarizers of different approaches, i.e., the GistSumm (of the shallow approach) and the RC-4 summarizer (of the deep approach). The *acronyms without explanation* problem had the greatest occurrence in the RSumm summarizer. In the MTRST-MCAD summarizer, 25.42% of the identified problems were related to *definite*

*noun phrase without reference to the previous mentions*, being the most frequent error for this summarizer.

Except for the *pronouns with misleading antecedents* problem, which was not identified in the summaries generated by MTRST-MCAD and RSumm systems, all the other errors happened at least in 1 summary of each summarizer. This shows that the summarizers did not treat or inadequately treated the problems that affect LQ.

TABLE 8 – Occurrence of each problem in the *corpus* per summarizer

| **Problems** | **Systems** | | | |
|---|---|---|---|---|
| | **MTRST-MCAD** | **GistSumm** | **Rsumm** | **RC-4** |
| RED | 5.46% | **30.71%** | 11.68% | **25.00%** |
| ACR-EXP | 12.83% | 18.62% | **27.92%** | 22.27% |
| DNP-REF | **25.42%** | 3.45% | 18.78% | 9.09% |
| SM+EXP | 5.23% | 16.51% | 6.60% | 14.09% |
| No_SEM | 19.95% | 4.22% | 8.63% | 5.91% |
| *Other* | 9.50% | 11.71% | 4.06% | 6.36% |
| 1M-EXP | 10.69% | 4.03% | 12.18% | 5.91% |
| CONTR | 0.95% | 4.80% | 1.02% | 4.55% |
| DM | 3.56% | 1.73% | 4.57% | 1.82% |
| PRO-ANT | 4.75% | 0.77% | 1.52% | 1.36% |
| INP+REF | 0.95% | 2.30% | 2.03% | 2.27% |
| INC_SENT | 0.71% | 0.96% | 1.02% | 0.45% |
| PRO_MIS | 0.00% | 0.19% | 0.00% | 0.91% |

Generally, the results of the annotation showed that the summarizers with the best summary informativeness evaluation in the area (RSumm and RC-4) also had a lower quantity of problems, but these summarizers still need to be improved, as there are LQ problems to be tackled.

It is interesting to notice that some of the error types in this work may be directly related to the classical ones of TAC. For example, the

*clause* category has problems such as*redundant information*, *connective/ discursive marker without appropriate context*, and *incomplete sentences*, which are directly related to *grammaticality* and to *no redundancy* in TAC.

Besides, the *no semantic relationship* problem of this category affects the *textual focus* of TAC, because a summary without semantic relationship among its sentences does not have a defined focus. The *referential clarity* of TAC is directly related to the *entity* category by means of the problems *definite noun phrase without reference to the previous mentions*, *indefinite noun phrase with reference to the previous mentions*, and *pronouns without antecedent*, for example. The *textual structure and coherence* errors are the merge of all errors that were considered.

The main problem observed in multi-document summaries in Friedrich *et al*. (2014) was *incomplete sentence*. On the other hand, the *redundant information* problem was the main problem in this work. However, these two problems are in the clause level, which may indicate that this is an important issue for future research.

In the experiments from Otterbacher *et al*. (2002), the *temporal ordering* problem was the most frequent one. This problem is related to the identification of correct temporal relationships between the events described in a summary. This problem is also at the clause level, which, in this work, happens in the *no semantic relationship* (No_SEM) error, when the temporal order of an event is not respected in the selection of sentences from the source texts to compose a summary.

### 6.2 Annotation agreement

As commented before in this paper, the annotation was made by a group, but we decided to measure the agreement among the annotators to check the understanding of the errors and the problem annotation process itself. For this, we calculated the Kappa measure and the percent agreement of the majority for 4 clusters of the CSTNews corpus (in particular, cluster C12, C22, C32 and C42). Notice that each cluster has 1 summary generated by each summarizer (GistSumm, RSumm, RC-4 and MTRST-MCAD), i.e., 4 summaries in each cluster.

Table 9 shows the Kappa scores for the agreement among annotators in each cluster for the simple indication of errors (in a binary decision).

TABLE 9 – Kappa measure for simply indicating a problem

| **Cluster** | **Kappa** |
|---|---|
| C12 | 0.409 |
| C22 | 0.641 |
| C32 | 0.578 |
| C42 | 0.324 |
| **Average** | **0.488** |

Cluster 22 had the best agreement result. However, due to the difficulty of the task, this result is not so high. The subjectivity causes different understandings and this is demonstrated when the annotators do the annotation in isolation. This behavior is repeated in Table 10, when we measure Kappa for the indication of the problem category. The Table 10 shows that the Kappa for the *Other* category had the best values. The agreement was the most significant in cluster 12 for the *Other* problem category.

TABLE 10 – Kappa measure for problem category

| **Cluster** | **Kappa for *Entity Level*** | **Kappa for *Clause Level*** | **Kappa for *Other*** |
|---|---|---|---|
| C12 | 0.356 | 0.560 | 1.000 |
| C22 | 0.670 | 0.537 | 0.902 |
| C32 | 0.552 | 0.616 | 0.627 |
| C42 | 0.606 | 0.418 | 0.751 |
| **Average** | **0.546** | **0.533** | **0.760** |

Considering the relatively low results of Kappa measure, the percent agreement by majority was also relevant in order to better judge the task. In this case, the percentage of sentences in all clusters that the majority of the annotators agreed was calculated. For example, in the summaries of cluster 12, at least 4 of the 6 annotators marked the occurrence of an error in all sentences (100% of the sentences, therefore)

of these summaries. Table 11 shows the results of the agreement by majority, considering the occurrence of a problem in a certain sentence.

Table 11 shows that the majority of the annotators agreed in marking an error in all the sentences in the summaries of clusters C12 and C22. Clusters C32 and C42 also presented a good percentage of agreement.

TABLE 11 – Percent agreement (by majority) for the indication of a problem in a sentence

| **Clusters** | **% of sentences** |
|---|---|
| C12 | 100 |
| C22 | 100 |
| C32 | 91.89 |
| C42 | 81.25 |
| **Average** | **93.28** |

We also used the agreement by majority for categories of problems. We calculated the percentage of sentences for which the majority of the annotators marked an error of a specific category. Table 12 shows the results obtained by this measure of agreement.

The majority of annotators agreed 100% for the sentences in the summaries of clusters C12 and C22, regarding the occurrence of all the problem categories. In cluster C42, the *clause* category was the only one in which the majority of the annotators agreed below 90%. These results showed that the majority of the annotators understood well all the linguistic problem categories identified in the summaries.

TABLE 12 – Percent agreement (by majority) for the indication of a problem category in a sentence

| **Clusters** | **% of sentences with problems** | | |
|---|---|---|---|
| | ***Entity Level*** | ***Clause Level*** | ***Other* problems** |
| C12 | 100 | 100 | 100 |
| C22 | 100 | 100 | 100 |
| C32 | 94.59 | 91.89 | 100 |
| C42 | 90.62 | 71.87 | 93.75 |
| **Average** | **96.30** | **90.94** | **98.43** |

To confirm this, Table 13 shows the percentage of sentences for which all annotators agreed in the identification of the linguistic problems.

TABLE 13 – Agreement for 100% of annotators for each problem

| **Problems** | **% of sentences** | | | |
|---|---|---|---|---|
| | **C12** | **C22** | **C32** | **C42** |
| 1M-EXP | 54.54 | 90.00 | 91.89 | 81.25 |
| SM+EXP | 81.81 | 76.66 | 84.48 | 90.62 |
| DNP-REF | 63.63 | 93.33 | 83.78 | 53.12 |
| INP+REF | - | - | 89.18 | - |
| PRO-ANT | - | - | - | 96.87 |
| ACR-EXP | 100 | 96.66 | 94.59 | 93.75 |
| No_SEM | 81.81 | 76.66 | 75.67 | 75.00 |
| DM | - | - | 91.89 | 96.87 |
| RED | - | 83.33 | 89.18 | 81.25 |
| CONTR | - | 86.66 | 94.59 | - |
| *Other* | - | 96.66 | 81.08 | 90.62 |

According to Table 13, over half of the sentences in the summaries had 100% of agreement among the annotators. All the sentences with the *acronyms without explanation* (ACR-EXP) problem were marked by all annotators for the first cluster. The hyphen (-) in Table 13 means that the error was not identified by any of the annotators. The *pronouns with misleading antecedents* (PRO_MIS) and *incomplete sentence* (INC_SENT) problems were not identified in the clusters used in the agreement and, for this reason, are not listed in the Table 13.

With the reported agreement results, we may conclude that the annotation task was well understood and the annotation is reliable. We believe that our well-defined typology of LQ problems was an important reason for the reported agreement scores.

## 7 Final remarks

This paper reported the study, an annotation task and the characterization of linguistic problems in multi-document summaries automatically produced by systems of varied paradigms, from shallow to deep approaches, including classic and state of the art methods. The corpus consisted of summaries composed by four automatic summarizers, and it was possible to verify that (i) some problems deserve more attention from the automatic summarizers, as problems related to redundancy and introduction of definite noun phrases and acronyms, which accounted for more than 50% of the errors, and (ii) that the summarizers with the best summary informativeness results (according to standard informativeness measures) also produce a lower quantity of problems. Our results may be used as a guide to treat errors in future summarizers.

The literature review and organization and the methodology used for the problem annotation process are also contributions to the area. In particular, the annotation strategy was interesting because the problem annotation involves difficult and fuzzy aspects as subjectivity and world knowledge, which may affect the consistency of the annotation. The agreement values confirmed that such annotation strategy is worthy following.

As future work, we consider to study error correlation in the summaries, as well as automatic methods for detecting and properly dealing with them, improving the summary quality.

For the interested reader, the corpus that was produced, the summarization systems that we used and other related information about this work may be found at the SUCINTO project webpage.[7]

**Acknowledgements**

The authors are grateful to FAPESP (*Fundação de Amparo à Pesquisa do Estado de São Paulo*), USP Research Office (PRP 668) and Federal University of Goiás for supporting this work.

**Contributions of the authors**

Márcio de Souza Dias (1rst author): responsible for organizing the corpus annotation task, writing the introductory sections of the paper and pulling together the sections written by the co-authors.
Ariani Di-Felippo (2nd author): responsible for describing and analyzing the linguistic problems annotated in the corpus of automatic summaries.
Amanda Pontes Rassi (3rd author): responsible for writing the description of the linguistic problems of the literature.
Paula Christina Figueira Cardoso (4th author): responsible for describing the summarization methods and the reference corpus in Portuguese for Automatic Summarization.
Fernando Antônio Asevedo Nóbrega (5th author): responsible for computing and describing all the annotation statistics.
Thiago Alexandre Salgueiro Pardo (6th author): responsible for describing the basic concepts of the Automatic Summarization domain.

## References

ANDO, R.; BOGURAEV, B.; BYRD, R.; NEFF, M. Multi-document Summarization by Visualizing Topical Content. *In*: ANLP/NAACL WORKSHOP ON AUTOMATIC SUMMARIZATION, 2000, New Brunswick. *Proceedings* […]. New Brunswick: Association for Computational Linguistics, 2000. p. 79-88. DOI: https://doi.org/10.3115/1117575.1117584

[7] Available on: http://conteudo.icmc.usp.br/pessoas/taspardo/sucinto/resources.html. Retrieved at: Feb. 10, 2019.

BEAUGRANDE, R.; DRESSLER, W. U. *Introduction to Text Linguistics.* 1. ed. London: Longman, 1981.

CARBONELL, J.; GENG, Y.; GOLDSTEIN, J. Automated Query-Relevant Summarization and Diversity-Based Reranking. *In*: IJCAI Workshop on AI in Digital Libraries, 1997, Nagoya. *Proceedings* […]. Nagoya: [s.n.], 1997. p. 12-19.

CARDOSO, P. C. F; MAZIERO, E.; JORGE, M.; SENO, E.; DI-FELIPPO, A.; RINO, L.; NUNES, M.; PARDO, T. A. S. CSTNews: A Discourse-Annotated Corpus for Single and Multi-Document Summarization of News Texts in Brazilian Portuguese. *In*: RST BRAZILIAN MEETING, 3., 2011, Cuiabá. *Proceedings* […]. Cuiabá: Sociedade Brasileira de Computação, 2011. p. 88-105.

CARDOSO, P. C. F.; PARDO, T. A. S. Joint Semantic Discourse Models for Automatic Multi-Document Summarization. *In*: BRAZILIAN SYMPOSIUM IN INFORMATION AND HUMAN LANGUAGE TECHNOLOGY, 10., 2015, Natal. *Proceedings* […]. Natal: Sociedade Brasileira de Computação, 2015. p. 81-90.

CARDOSO, P. C. F.; PARDO, T. A. S. Multi-Document Summarization Using Semantic Discourse Models. *Procesamiento de Lenguaje Natural,* Jaén, Espanha, v. 56, n. 1, p. 57-64, 2016.

CARLETTA, J. Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics*, Cambridge, v. 22, n. 2, p. 249-254, 1996.

CASTRO JORGE, M. L. R. *Modelagem gerativa para sumarização automática multidocumento*. 2015. 151f. Tese (Doutorado em Ciência de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2015.

CASTRO JORGE, M. L. R.; PARDO, T. A. S. Experiments with CST-based Multidocument Summarization. *In*: ACL WORKSHOP: GRAPH-BASED METHODS FOR NATURAL LANGUAGE PROCESSING, 5., 2010, Uppsala, Sweden. *Proceedings of TextGraphs-5* […]. Uppsala: Association for Computational Linguistics, 2010. p. 74-82.

CONROY, J. M.; SCHLESINGER, J. D.; KUBINA, J.; RANKEL, P. A.; O'LEARY, D. P. CLASSY 2011 at TAC: Guided and Multilingual Summaries and Evaluation Metrics. *In*: TEXT ANALYSIS

CONFERENCE, 4., 2011, Maryland. *Proceedings* […]. Maryland: NIST, 2011. p. 1-8.

CRISTINI, L. F.; DI-FELIPPO, A. Violações linguísticas em referências a entidades do tipo "pessoa" em extratos automáticos multidocumento. *In*: WORKSHOP ON PORTUGUESE DESCRIPTION, 6., 2019, Salvador. *Proceedings* […]. Salvador: [s.n], 2019. p. 244-252.

DANG, H. T. Overview of DUC 2005. *In*: DOCUMENT UNDERSTANDING CONFERENCE, 2005, Vancouver. *Proceedings* […]. Vancouver: NIST, 2005 p. 1-12. Available on: https://duc.nist.gov/pubs.html#2005. Retrieved at: January. 2015.

FONSECA, H. P. A.; DIAS, M. S.; SILVA, N. F. F. Identificação automática de erros em sumários multidocumento. *In*: SYMPOSIUM IN INFORMATION AND HUMAN LANGUAGE TECHNOLOGY, 12., 2019, Salvador. *Anais*... Salvador: Brazilian Computer Society, 2019. p. 395-399.

FRIEDRICH, A.; VALEEVA, M.; PALMER, A. LQVSumm: A Corpus of Linguistic Quality Violations in Multi-Document Summarization. *In*: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION, 9., 2014, Reykjavik. *Proceedings* […]. Reykjavik: European Language Resources Association, 2014. p. 1591-1599.

GIANNAKOPOULOS, G.; KARKALETSIS, V. AutoSummENG and MeMoG in Evaluating Guided Summaries. *In*: TEXT ANALYSIS CONFERENCE, 4., 2011, Maryland. *Proceedings* […]. Maryland: NIST, 2011. p. 1-10.

HAGHIGHI, A.; VANDERWENDE, L. Exploring Content Models for Multi-Document Summarization. *In*: HUMAN LANGUAGE TECHNOLOGIES: THE ANNUAL CONFERENCE OF THE NORTH AMERICAN CHAPTER OF THE ACL, 2009, Boulder. *Proceedings* […]. Boulder: NACL, 2009. p. 362-370. DOI: https://doi.org/10.3115/1620754.1620807

HOVY, E. H.; LAVID, J. M. Towards a Science of Corpus Annotation: A New Methodological Challenge for Corpus Linguistics. *International Journal of Translation Studies,* [*S.l.*], v. 22, n. 1, p. 13-36, 2010.

KASPERSSON, T.; SMITH, C.; DANIELSSON, H.; JÖNSSON, A. This Also Affects the Context – Errors in Extraction Based Summaries. *In*: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION, 8., 2012, Istanbul. *Proceedings* […]. Istanbul: European Language Resources Association, 2012. p. 173-178.

KOCH, I. G. V. *A coesão textual*. 10. ed. São Paulo: Contexto, 1998.

KOCH, I. G. V.; TRAVAGLIA, L. C. *A coerência textual*. São Paulo: Contexto, 2002.

LIN, C-Y. ROUGE: A Package for Automatic Evaluation of Summaries. *In*: ACL WORKSHOP ON TEXT SUMMARIZATION BRANCHES OUT, 2004, Barcelona. *Proceedings* […]. Barcelona: ACL, 2004. p. 74-81.

LIN, Z.; LIU, C.; NG, H. T.; KAN, M. Combining coherence models and machine translation evaluation metrics for summarization evaluation. In: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 50., 2012, Jeju Island. *Proceedings* […]. Jeju Island: ACL, 2012. p. 1006-1014.

MANI, I. *Automatic Summarization*. Amsterdam: John Benjamins Publishing, 2001.

MANI, I.; MAYBURY, M. T. *Advances in Automatic Text Summarization.* Cambridge: The MIT Press. 1999. DOI: https://doi.org/10.1075/nlp.3

MANN, W. C.; THOMPSON, S. A. Rhetorical Structure Theory: A Theory of Text Organization. *Technical Report ISI/RS-87-190*, 1987. Available on: https://www.sfu.ca/rst/05bibliographies/bibs/ISI_RS_87_190.pdf. Retrieved at: March. 2015.

MARCU, D. Discourse Trees Are Good Indicators of Importance in Text. *In*: MANI, I.; MAYBURY, M. T. (ed.). *Advances in Automatic Text Summarization*. Cambridge: The MIT Press, 1999. 123-136.

MCKEOWN, K.; RADEV, D. R. Generating Summaries of Multiple News Articles. *In*: ANNUAL INTERNATIONAL ACM-SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 18., 1995, Seattle. *Proceedings* […]. Seatle: Association for Computing Machinery, 1995. p. 74-82. DOI: https://doi.org/10.1145/215206.215334

MIHALCEA, R.; TARAU, P. An Algorithm for Language Independent Single and Multiple Document Summarization. *In*: INTERNATIONAL JOINT CONFERENCE ON NATURAL LANGUAGE PROCESSING, 2., 2005, Jeju Island. *Proceedings* […]. Jeju Island: ACL, 2005. p. 19-24. DOI: https://doi.org/10.1007/11562214

NENKOVA, A.; MCKEOWN, K. R. *Automatic Summarization*. Foundations and Trends in Information Retrieval. Hanover, MA: Now Publishers, 2011. DOI: https://doi.org/10.1561/1500000015

OLIVEIRA, P. C. F. de. CatolicaSC at TAC 2011. *In*: TEXT ANALYSIS CONFERENCE (TAC), 4., 2011, Gaithersburg. *Proceedings* […]. Gaithersburg: NIST, 2011. p. 1-3.

OTTERBACHER, J. C.; RADEV, D. R.; LUO, A. Revisions that Improve Cohesion in Multi-Document Summaries: A Preliminary Study. *In*: ACL-02 WORKSHOP ON AUTOMATIC SUMMARIZATION, 2002, Philadelphia. *Proceedings* […]. Philadelphia: ACL, 2002. p. 27-36. DOI: https://doi.org/10.3115/1118162.1118166

OWCZARZAK, K.; DANG T. H. Overview of the TAC 2011 Summarization Track: Guided task and AESOP task. *In*: TEXT ANALYSIS CONFERENCE, 3., 2011, Gaithersburg. *Proceedings* […]. Gaithersburg: NIST, 2010. Available on: https://tac.nist.gov/2011/Summarization/Guided-Summ.2011.guidelines.html. Retrieved at: January. 2015.

PARDO, T. A. S.; RINO, L. H. M.; NUNES, M. G. V. GistSumm: A Summarization Tool Based on a New Extractive Method. *In*: WORKSHOP ON COMPUTATIONAL PROCESSING OF THE PORTUGUESE LANGUAGE, 6., 2003, Faro, Portugal. *Proceedings* […]. Faro: Springer, 2003. p. 210-218. DOI: https://doi.org/10.1007/3-540-45011-4_34

PARDO, T. A. S. GistSumm - GIST SUMMarizer: extensões e novas funcionalidades. *Technical Report NILC-TR-05-05,* 2005. Available on: https://sites.icmc.usp.br/taspardo/NILCTR0505-Pardo.pdf. Retrieved at: January. 2015.

PITLER, E.; LOUIS, A.; NENKOVA, A. Automatic Evaluation of Linguistic Quality in Multi-document Summarization. *In*: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 48., 2010, Uppsala, Sweden. *Proceedings* […]. Uppsala: ACL, 2010. p. 544-554.

RADEV, D. R. A Common Theory of Information Fusion from Multiple Text Sources, Step One: Cross-document Structure. *In*: ACL SIGDIAL WORKSHOP ON DISCOURSE AND DIALOGUE, 1., 2000, Hong Kong. *Proceedings* […]. Hong Kong: ACL, 2000. p. 74-83. DOI: https://doi.org/10.3115/1117736.1117745

RADEV, D. R.; TEUFEL, S.; SAGGION, H.; LAM, W.; BLITZER, J.; CELEBI, A.; QI, H.; LIU, D.; DRABEK, E. Evaluation Challenges in Large-Scale Multi-Document Summarization. *In*: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 41., 2003, Sapporo, Japan. Proceedings […]. Sapporo: ACL, 2003. p. 375-382. DOI: https://doi.org/10.3115/1075096.1075144

RIBALDO, R.; AKABANE, A. T.; RINO, L. H. M.; PARDO, T. A. S. Graph-based Methods for Multi-Document Summarization: Exploring Relationship Maps. Complex Networks and Discourse Information. *In*: INTERNATIONAL CONFERENCE ON COMPUTATIONAL PROCESSING OF PORTUGUESE, 10., 2012, Coimbra. *Proceedings* (Lecture Notes in Computer Science 7243) […]. Coimbra: Springer, 2012. p. 260-271. DOI: https://doi.org/10.1007/978-3-642-28885-2_30

RIBALDO, R.; CARDOSO, P. C. F.; PARDO, T. A. S. Exploring the Subtopic-Based Relationship Map Strategy for Multi-Document Summarization. *Journal of Theoretical and Applied Computing* (RITA), Porto Alegre, RS, v. 23, n. 1, p. 183-211, 2016. DOI: https://doi.org/10.22456/2175-2745.59104

SALTON, G.; SINGHAL, A.; MITRA, M.; BUCKLEY, C. Automatic Text Structuring and Summarization. *Information Processing & Management*, [*S.l.*], v. 33, n. 2, p. 193-207, 1997. DOI: https://doi.org/10.1016/S0306-4573(96)00062-3

ZHANG, Z.; GOLDENSHON, S. B.; RADEV, D. R. Towards CST-Enhanced Summarization. *In*: NATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE, 18., 2002, Menlo Park, CA. *Proceedings* […]. Menlo Park: AAAI, 2002. p. 439-445.