Multi-Document Summarization: Content Selection based on CST Model (Cross-document Structure Theory)

Maria Lucia Castro Jorge[1] and Thiago Alexandre Salgueiro Pardo[1]

[1] Núcleo Interinstitucional de Linguística Computacional (NILC)
Instituto de Ciências Matemáticas e de Computação- Universidade de São Paulo
Av. Trabalhador São Carlense, 400- Centro
Caixa Postal: 668 – CEP: 13560-970 - São Carlos-SP
{mluciacj, taspardo}@icmc.usp.br

Abstract. This paper presents the definition, formalization and evaluation of Content Selection strategies, based on CST semantic-discursive model (Cross-document Structure Theory). These strategies were modeled by operators that represent possible summarization preferences. Our experiments were performed using CSTNews corpus, which consists on a group of journalistic texts written in Brazilian Portuguese language, and show that the use of CST knowledge improves the quality of summaries in terms of informativity. This approach is novel for Brazilian Portuguese language, because it is the first that explores linguistic knowledge in a differentiated way.

Keywords: Multi-document, Summarization, CST, Content Selection.

1 Introduction

The increasing of new technologies has had an impact on the amount of available information, especially on-line. Much of this information is redundant, complementary and contradictory, because it comes from different sources. Consequently, Multi-document Summarization appears to be a useful resource.

Multi-Document Summarization (MDS) is the automatic production of a unique summary from a group of texts on a same topic or related topics [5],[6]. MDS has very important challenges such as capturing the most important information of a topic within a generic perspective or prioritizing information preferences specified by the user. It is also important to maintain coherence and cohesion in the summary by treating redundancy and organizing information properly among other challenges.

In this work, we focus on the Content Selection task. We assume that it is done a previous representation of source texts according to CST model [11], which provides information of semantic-discursive nature. We explore different Content Selection strategies for different summarization preferences that satisfy the user's particular requirements. After Content Selection, we simply perform the juxtaposition of the selected content, producing the final summary. The hypothesis of this work is that linguistic knowledge provided by CST will help producing better summaries in terms of quality and informativity.

The proposed strategies are formalized in operators for content selection, containing rules for manipulation of the content. Our experiments were performed using a journalistic corpus of texts written in Portuguese, CSTNews [2], and they show that the usage of CST knowledge helps improving the informativity in summaries, confirming our hypothesis.

This work has various contributions, the most important ones are: being the first investigation that proposes semantic-discursive methods for MDS of Brazilian Portuguese texts; outperforming summarizers for texts written in Portuguese in terms of informativity; validating and refining the CST model, making it more consistent.

Next, in Section 2, the main related works will be studied. In Section 3, the Content Selection operators will be presented. In Section 4, the evaluation results will be discussed. Finally, in Section 5 some final remarks will be given.

2 Related Work

CST model was originally proposed as a set of 24 relations that make explicit the relations among different parts of the texts. These relations represent common multi-document phenomena, such as redundancy, complementarity and contradiction among information units, explored in detail by [3]. To illustrate how these relations occur among texts, in Table 1 there are shown some parts of journalistic texts related to a common topic. The examples where extracted from the corpus.

Table 1. Examples of CST relations

Sentence 1	Sentence 2	CST Relation
Ao menos 17 pessoas morreram após a queda de um avião de passageiros na República Democrática do Congo.	Um acidente aéreo na localidade de Bukavu, no leste da República Democrática do Congo (RDC), matou 17 pessoas na quinta-feira à tarde, informou nesta sexta-feira um porta-voz das Nações Unidas.	subsumtion (←) Sentence2 subsumes Sentence 1

It is important to notice that relations may have directionality. In the example, the subsumption relation goes from Sentence 2 to Sentence 1 (from right to left, which is represented by ←), since Sentence 2 contains all the information of Sentence 1 and provides extra information. On the other hand, other relations such as contradiction don’t have directionality, since sentences contradict each other in the same way.

As part of this work, the set of CST relations were refined in order to obtain a better formalization and improving annotation concordance, by reducing ambiguity. More details on this refinement can be found in [3].

Some works have employed CST for MDS, the first one was [11] who proposed a 4-stage methodology for summarization: text clustering, internal structuring of the texts, establishment of CST relations and sentence selection. Sentences are ranked according to their importance and then selected.

[8] propose a methodology consisting of an information extraction system, which extracts information and organizes it in templates (previously defined structures in which information is organized). Next, the semantic relations (similar to CST) are established among templates. According to the relations, templates can be combined or eliminated in order to build the final summary.

Other important works are: [14], who proposed CST based re-ranking methods for superficial summarizers rankings; [9] investigated how CST relations can help to improve cohesion in summaries, by grouping closely sentences linked by CST relations; [1] proposed a new classification of semantic-discursive relations, dividing the relations into two categories: synchronous (a fact or event is described by different sources at a particular moment) and diachronic (the evolution of facts or events in a period of time). Then, the authors proposed a summarization methodology based on previously defined templates and ontologies. Their work was just theoretical.

In the next section, our methodology will be described.

3 Definition and Formalization of Content Selection Operators

We define a Content Selection operator as a computational artifact that processes a content representation and produces a shorter version that contains the most relevant information, according to some previously specified criteria. In this work, a content representation corresponds to the CST analysis of the texts. Therefore, the operators are applied after the source texts have already been analyzed according to CST model. Currently, this analysis has to be done manually, since the first automatic parser for Portuguese is still under development [7].

The input for the operators is an initial rank of sentences extracted from the source texts. This rank is obtained from the graph constructed in the CST analysis (in the rest of the paper we will refer to this graph as CST graph), where sentences are nodes and CST relations are the edges of the graph. The more relevant a sentence is, the more on top positions of the rank it should be located. The function of the operators is to produce a refined rank, so that relevant sentences, according to a given preference, achieve higher positions in the rank. Preferences reflect possible particular requirements that a user might find important to include in a summary such as context information or the evolution of an event in time. Finally, given the compression rate, a number of sentences from the refined rank are selected to compose the final summary.

In the initial rank, the relevance of sentences depends on the number of CST relations the sentence has, this is, sentences with more CST relations are considered more relevant. In Fig.1, it is shown an hypothetical example of a CST graph and the initial rank built from it.

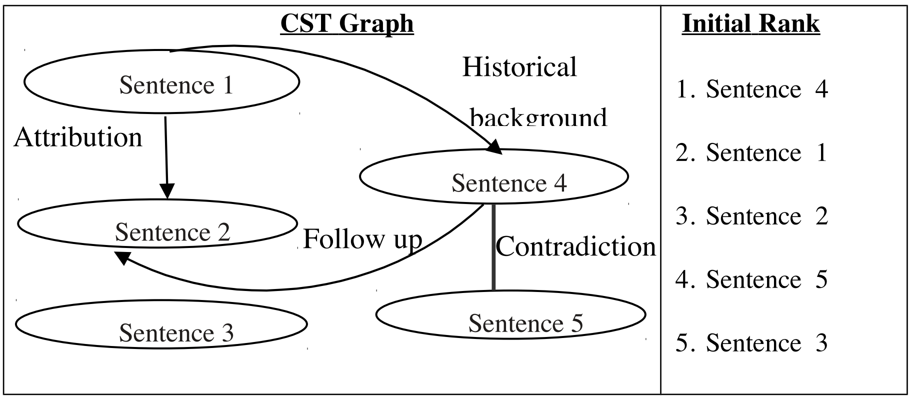

Fig. 1. Example of Initial Rank and it's correspondent CST Graph

CST based content selection operators are defined in template form containing a set of rules. These rules are specified in terms of conditions and restrictions, that, if satisfied, they will apply some functions in order to modify the initial rank. Each rule is defined by: CONDITIONS, RESTRICTIONS→ACTIONS. Each condition has the format CONDITION (Si, Sj, Direction, Relation), and is satisfied if exists the specified relation and direction (from Si to Sj: →; or the opposite case ←; or no direction at all —) between two sentences Si and Sj. The restrictions are optional, since they represent possible extra requirements for the operator to be applied.

If all the conditions and restrictions are satisfied, then the actions are applied to the initial rank, which will produce a refined version of the rank. Actions are defined in terms of, at least, one of the following functions:

– **GO_UP** (S_i,S_j): Sentence j is put on the immediate after position of sentence i in the rank; it is important to notice that sentence i will always be above to sentence j in the rank.
– **SWITCH**(S_i,S_j): Sentences in the positions i and j are switched.
– **DELETE** (S_j): Sentence j is deleted from the rank.

For this work, we define operators for possible summarization strategies (which represent summarization preferences) of Content Selection. These strategies are: Presentation of Context Information, which gives priority to context information such as sentences with elaboration or historical background relations; Presentation of Contradictory Information, which gives priority to sentences with contradiction relation; Identification of Authorship, which gives priority to sentences with attribution and citation relations; Redundancy Treatment, which eliminates redundancy by exploring relations such as identity, subsumption and equivalence; and finally, Presentation of Events Evolution in time, which gives priority to sentences with follow-up relation. The process of constructing the initial rank can also be specified in the form of an operator, in which the preference is the main information of the topic in a generic context (just sentences with more CST relations). We call this operator as "Generic Operator" or "Operator for Main Information Retrieval".

Each operator is defined by three fields: a reference name, a brief description of the purpose of the operator, and the set of rules. To illustrate this, in Table 2 it is shown the definition of the operator for presentation of context information. In this operator it is searched for sentences (along the rank) that present relations of the type historical background and elaboration, since those relations are the ones that provide context information.

Table 2. Operator for presentation of Context Information

Name	Operator for presentation of Context Information
Description	Preference on historical and complement information
Regras	CONDITION (S_i, S_j, ←, Elaboration) ⇒GO_UP (S_i, S_j) CONDITION (S_i, S_j, ←, Historical background) ⇒ GO_UP (S_i, S_j)

Each rule is verified for all sentences of the initial rank. This verification is performed in the order that the rules are read from the template. Different orders of application can lead to different results. In this work it has not been deeply investigated the best order of application of the rules, though we consider that the first rules to be applied are the ones dealing with relations that tend to appear more in the corpus, so that rules dealing with relations that appear less in the corpus will not cause many re rankings. The application of the operator described in Table 2 to the initial rank in Fig. 1 will produce a refined rank, in which Sentence 1 will go up in the rank, above Sentence 2 and after Sentence 4, since it presents a historical background relation with Sentence 4. This is shown in Fig. 2.

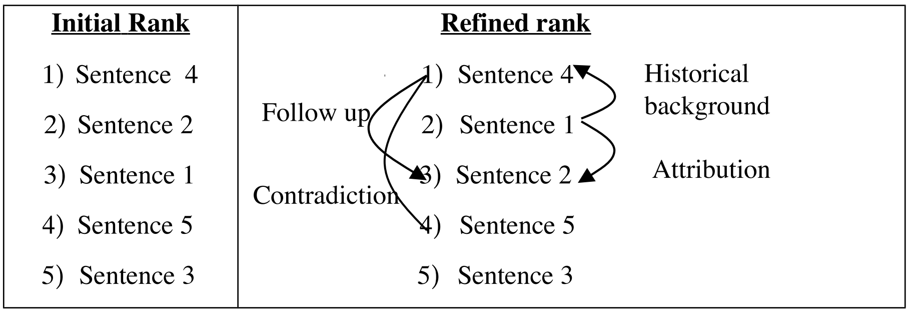

Fig. 2. Refined Rank after operator is applied

After the refinement of the rank, some sentences may still be kept redundant, in order to solve this problem, it is necessary to apply the operator for redundancy treatment, which deletes from the rank the sentences that have redundant content (indicated by relations such as equivalence, identity, etc.). According to compression rate, a number of sentences will be indicated by the refined rank to build the final summary. Considering the example of Fig. 2 and a compression rate that only allows the inclusion of 2 sentences in the summary, sentences 4 and 1 would be chosen.

Not all CST relations can be properly treated. For example, when overlap relation occurs, we cannot simply choose to delete one of the sentences, because both sentences introduce novel information. For this work, we have treated this case by applying sentence fusion proposed by [13]. In the next Section, the results of the evaluation are explained and discussed.

4 Experiments and Results

In order to evaluate the content selection operators, we built a multi-document summarizer prototype which we call CSTSumm (CST Summarizer). This prototype applies the procedures explained in Section 3.

For our experiments, we used the corpus CSTNews, which is composed of 50 clusters of journalistic texts written in Brazilian Portuguese language. Each cluster contains 2 or 3 documents on a same topic. For this corpus, it was done manually a CST analysis of the texts and, also, there were produced generic human summaries (extracts and abstracts) for each correspondent cluster. The size of the summaries corresponds to the 30% of the size of the biggest text of the cluster (considering that the size is given in terms of the number of words).

The automatic summaries were produced considering the same compression rate of the human summaries. In this work, it is considered two evaluation methods: automatic evaluation, which is used to evaluate the informativity of generic summaries, and human evaluation, which is used to evaluate more subjective factors such as: coherence, cohesion, grammar correctness and redundancy in summaries.

For the automatic evaluation it was used the ROUGE measure [4]. This measure produces result values in terms of f-measure, precision and recall. The results for our methods were compared with the summarizer GistSumm [10], which is the only summarizer available for texts written in Portuguese, and also with MEAD [12], which is a well-known multilingual summarizer. Moreover, we extended our experiments by applying the methodology proposed by [14], using the two summarizers mentioned above, which are considered to be superficial summarizers, since they don't make any use of linguistic knowledge. In particular, the sentences of the ranks produced by these summarizers were re ranked according to CST relations.
In Table 3, it is shown the results resume for the automatic evaluation. It can be observed that summaries produced by operators have better results than GistSumm and MEAD, in terms of f-measure. Also, the performance of MEAD and GistSumm improves when CST relations are taken into account, as predicted by [14]. On the other hand, human evaluation results showed a good performance considering the factors mentioned above. According to this evaluation, informativity was one the

factors with better punctuation. It was also considered that summaries had a low presence of redundant information.

Table 3. ROUGE results

	PRECISION	RECALL	F-MEASURE
Main Information Retrieval	0.5564	0.5303	0.5356
Redundancy Treatment	0.5761	0.5065	0.5297
Presentation of Contradictions	0.5503	0.5379	0.5365
Authorship Identification	0.5563	0.5224	0.5310
Presentation of events' evolution in time	0.5159	0.5222	0.5140
Presentation of Context Information	0.5196	0.4938	0.4994
GistSumm	0.3599	0.6643	0.4599
GistSumm using CST	0.4945	0.5089	0.49940
MEAD	0.5242	0.4602	0.4869
MEAD using CST	0.5599	0.4989	0.5230

On the other hand, human evaluation results showed a good performance. According to this evaluation, informativity and redundancy elimination was one the factors with better punctuation. In Fig. 3 there are shown two summaries produced by CSTSumm. The first summary was built applying the Generic operator and the second summary was build applying the Context Information operator.

Summary 1

Um terremoto de 6,8 graus na escala Richter atingiu a costa noroeste do Japão nesta segunda-feira, 16, matando pelo menos sete pessoas na cidade de Kashiwazaki e deixando outros 700 feridos. O terremoto, que pôde ser sentido em Tóquio, foi seguido por outro tremor de menor magnitude, de 4,2 graus na escala Ritcher, às 10h34m. Chamas e rolos de fumaça preta foram vistos na usina nuclear de Kashiwazaki, que foi automaticamente fechada durante o terremoto.

Summary 2

Chamas e rolos de fumaça preta foram vistos na usina nuclear de Kashiwazaki, que foi automaticamente fechada durante o terremoto. Os prédios chegaram a tremer em Tóquio, e os reatores de usinas nucleares em Niigata desligaram-se automaticamente para checagens, embora não haja relatos de vazamento de radiação. O Japão é um dos países do mundo mais suscetíveis a terremotos, com um tremor ocorrendo a ao menos cada cinco minutos.

Fig. 3. CSTSumm summaries

5 Final Remarks

In this work we showed that CST model allows exploring the knowledge among texts on the same topic, which helps to select content that improves informativity in summaries. For the moment, our system only applies one operator at a time, except for the redundancy operator that can be applied together with the other operators. For future works, we plan to study techniques that allow the application of various operators at a time.

Acknowledgments. The authors are thankful to CAPES, CNPq and FAPESP for their financial support.

References

1. Afantenos, S.D., Doura, I., Kapellou, E., Karkaletsis, V.: Exploiting Cross-Document Relations for Multi-document Evolving Summarization. In: Proceedings of SETN, pp. 410-419. (2004) .
2. Cardoso, P.C.F., Maziero, E.G., Jorge, M.L.C., Seno, E.M.R., Di Felippo, A., Rino, L.H.M., Nunes, M.G.V., Pardo, T.A.S.: CSTNews - A Discourse-Annotated Corpus for Single and Multi-Document Summarization of News Texts in Brazilian Portuguese. In: Proceedings of the 3rd RST Brazilian Meeting, pp. 88-105. Cuiabá/MT, Brazil . (2011).
3. Castro Jorge, M.L.R.: Sumarização Automática Multidocumento: seleção de conteúdo com base no Modelo CST (Cross-document Structure Theory). Dissertação de Mestrado. ICMC-Universidade de São Paulo. São Carlos-SP, Abril, 86p. (2010).
4. Lin, C.Y., Hovy, E.: Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics. In: Proceedings of 2003 Language Technology Conference. Edmonton, Canada. (2003).
5. Mani, I.: Automatic Summarization. John Benjamins Publishing Co. Amsterdam. (2001).
6. Mani, I., Maybury, M. T.: Advances in automatic text summarization. MIT Press, Cambridge/MA. (1999).
7. Maziero, E.G., Jorge, M.L.C., Pardo, T.A.S.: Identifying Multidocument Relations. In: Proceedings of the 7th International Workshop on Natural Language Processing and Cognitive Science - NLPCS, pp.60-69. Funchal/Madeira, Portugal. (2010).
8. McKeown, K., Radev, D.R.: Generating summaries of multiple news articles. In: Proceedings of the 18th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, pp. 74-82, Seattle,WA.(1995).
9. Otterbacher, J.C., Radev, D.R., Luo, A.: Revisions that improve cohesion in multi-document summaries: a preliminary study. In: Proceedings of the Workshop on Automatic Summarization, pp 27-36. (2002).
10. Pardo, T.A.S.: GistSumm - GIST SUMMarizer: Extensões e Novas Funcionalidades. Série de Relatórios do NILC. NILC-TR-05-05. São Carlos,Brasil. (2005).
11. Radev, D.R.: A common theory of information fusion from multiple text sources, step one: Cross-document structure. In: Proceedings of the 1st ACL SIGDIAL Workshop on Discourse and Dialogue, pp 74-83. Hong-Kong, China. (2000).
12. Radev, D.R., Blair-Goldensohn, S., Zhang, Z.:Experiments in single and multi-document summarization using MEAD. In: Proceedings of the First Document Understanding Conference. New Orleans/LA. (2001).
13. Seno, E.R.M., Nunes, M.G.V.: Reconhecimento de Informações Comuns para a Fusão de Sentenças Comparáveis do Português. Linguamática, Vol. 1, pp. 71-87. (2009).
14. Zhang, Z., Goldenshon, S.B., Radev, D.R.: Towards CST-Enhanced Sumarization. In: Proceedings of the 18th National Conference on Artificial Intelligence. (2002).