2009 Seventh Brazilian Symposium in Information and Human Language Technology

Content Selection Operators for Multidocument Summarization based on Cross-document Structure Theory

Maria Lucía Castro Jorge, Thiago Alexandre Salgueiro Pardo

Núcleo Interinstitucional de Lingüística 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 aims at presenting an analysis of content selection techniques for multidocument summarization based on the multidocument discourse theory CST (Cross-document Structure Theory). We approach the task of content selection by using CST-based operators and focus specifically on redundancy treatment, which is an important and pervasive problem in multidocument summarization. Our experiments with Brazilian Portuguese news texts show that CST improves summaries quality by exploring relations among texts. Particularly, redundancy is reduced by identifying common information among texts, especially when compression rate is low.*

1. Introduction

Over the last decade, new technologies have led to an incredible increase of information amount. Consequently, processing this information has become a more difficult task. Lots of topics are widely spread in different on-line sources. Some sources report evolving events; other sources repeat some information. Within this scenario, Multidocument Summarization (MDS) may be helpful. It consists in producing a unique summary from a group of texts about the same topic or of related topics (Mani, 2001).

Content selection is the summarization step in which the relevant information that will be in the summary is selected (Mani and Maybury, 1999). A multidocument summary must contain at least the most relevant information from the texts, but there are also some other challenges in MDS that have to be addressed. Since several texts are dealing with the same topic, there will be common, contradictory and complementary information to deal with.

Basically, there are two approaches for content selection: the superficial one, which use statistical metrics for selecting the information to be in the summary, and the deep one, which makes use of linguistic and computational-linguistic knowledge for performing the selection. Hybrid approaches are also possible. In this paper, we focus on the deep approach, specifically on the use of CST (Cross-document Structure Theory) (Radev, 2000).

Radev proposes CST as a way of exploring groups of texts with related content by establishing relations between their parts. These relations explore similarities and differences between the content of texts units and, therefore, are useful for better understanding and dealing with textual information, mainly for multidocument processing. For MDS, CST may be helpful in identifying the relatedness of information units for producing better summaries.

Within this context, the main goal of this paper is to explore content selection techniques for MDS based on CST, focusing specifically on redundancy treatment, since redundancy is an important and pervasive problem in MDS. Based on previous work in the area, we approach the problem by developing content selection operators based on CST. We run some experiments for Brazilian Portuguese news texts and show that CST-based methods help reducing redundancy in automatic summaries, especially when compression rate is low.

978-0-7695-3945-4/09 $26.00 © 2009 IEEE
DOI 10.1109/STIL.2009.15

Summary informativeness also improves when content that is highly CST-related is selected. It is important to say that this work builds on the work presented by Aleixo and Pardo (2008a).

This paper is organized as follows. In Section 2, CST and related works on MDS are briefly presented. Our content selection methodology is presented in Section 3. Some experiments and their results are reported in Section 4. We present some final remarks in Section 5.

2. Related Work

2.1. Cross-document Structure Theory (CST)

Inspired by Rhetorical Structure Theory (Mann and Thompson, 1987) and on the works of Trigg (1983) and Trigg and Weiser (1987), CST appears as a theory/model for relating multiple texts on related topics of general domain.

For purposes of MDS, a group of texts has certain properties that must be treated, for example, it may be found statements of agreement, contradictions, and complementary information, which may be modeled by CST. In Figure 1, there is an illustration of two related texts that present the mentioned properties. One may see redundant (the airplane accident), complementary (Bukavu is in the east of Democratic Republic of Congo), and contradictory information (13 vs. 17 victims).

An airplane accident in Bukavu, east of Democratic Republic of Congo, killed 13 people this Thursday in the afternoon, informed last Friday an employee of the ONU.
At least 17 people died after an airplane fell down at Democratic Republic of Congo. An ONU employee said that the airplane, of Russian fabrication, was trying to land at Bukavu's airport in the middle of a storm.

Figure 1. Example of MDS scenario

The set of 24 relations that CST originally proposes represent well these phenomena. These 24 relations are listed in Table 1. For instance, using these relations, one might establish a contradiction relation among the first and second sentences of Figure 1, as well as an overlap relation among them.

Table 1. CST relations proposed by Radev (2000)

Identity	*Modality*	*Judgment*	*Cross-Reference*
Equivalence (paraphrasing)	*Attribution*	*Fulfilment*	*Citation*
Translation	*Summary*	*Description*	*Refinement*
Subsumption	*Follow-up*	*Reader profile*	*Agreement*
Contradiction	*Elaboration*	*Contrast*	*Generalization*
Historical background	*Indirect-Speech*	*Parallel*	*Change of perspective*

CST also has a general schema in which relations among texts units of different granularities are represented. In Figure 2, there is an illustration of CST general schema (CST graph). The figure is reproduced exactly as it appears in the work of Radev (2000, p. 78). As it can be seen, words, sentences, phrases or even the whole documents/texts may be considered as text units. CST relations can be established at any level of analysis. According to the theory, only a subset of the text units should be related, because, in general, there may be parts in the texts that do not refer to the same subject. The established relations may also have directionality, while others may not. For example, the "*equivalence*" relation has no directionality, since both text units it relates have the same content. On the other hand, "*historical background*" relation has directionality, since one text unit is giving a historical context to the other one that is related to the first one.

In order to determine which text units should be related, a lexical similarity measure must be applied before initiating the process of establishing CST relationships across

documents. This fact reduces the number of text unit pair combinations, otherwise, taking the whole set of combinations would be a high cost task, as Zhang et al. (2002) argue. In general, CST analysis is ambiguous like any other subjective analysis, since different human annotators may identify different relations between the same text units and, therefore, agreement between humans is low. Because of this, CST has been criticized in other works like (Afantenos, 2004) and (Zhang et al., 2002).

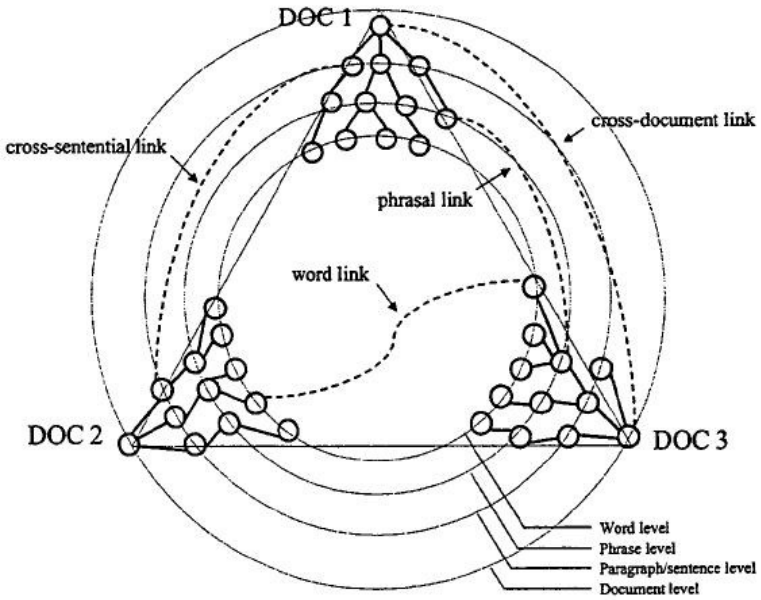

Figure 2. CST general schema (Radev, 2000, p. 78)

In this paper, following Zhang et al. proposal, we work with a refinement of the original CST relations, in which we consider only 14 relations instead of the whole set of 24 relations. This refinement was made by joining some related relations (e.g., description and elaboration relations) in order to reduce ambiguity and to improve agreement between human annotators, as well as by eliminating some relations that would never appear in our scenario (e.g., reader profile and change of perspective relations). In annotation experiments for Brazilian Portuguese news texts, the refined theory achieved a good agreement for 4 human annotators for both relation type and directionality. Previous initiatives for Portuguese were not able to get good results, as reported by Aleixo and Pardo (2008b).

2.2. Multidocument Summarization with Cross-document Structure Theory

Several investigations have used CST for MDS, including its author. Radev (2000) not only proposed CST, but also proposed a four stage MDS methodology. In the first stage, documents should be clustered according to content similarity; in the second stage, an internal structuring should be made for each document, possibly involving lexical, syntactic and semantic structuring; in the third stage, CST relations should be established across documents and information units be organized as a graph in which each node represents a text unit and connections represent CST relations; finally, in the fourth stage, text units should be selected according to CST relations in order to compose the final summary. For this last stage, Radev proposes the creation of operators encoding user preferences, which select content across the CST graph by using the knowledge within each relation. A particular example is the redundancy operator, which explores the graph of relations, selects relevant information and discards redundant information. In this case, relations like subsumption, equivalence, identity, and overlap should help excluding repeated information in the final summary. Operators might also express other user preferences, for example, a user may want a summary containing

information written by the same author, information from a particular source, or information containing contextual information. Radev also proposes a general and simple operator in which there is no particular preference considered. In this case, units that have more CST connections in the graph should represent important content. Radev proposal is based on previous work of Radev and McKeown (1998), where the idea of operators was born (although their operators were not explicitly CST-based operators).

Another important work based on CST is the one of Zhang et al. (2002). The authors aim at improving the quality of rank-based summarizers by using CST relations. They re-rank sentences according to CST information. For example, sentences that were ranked last according to statistical measures may be re-ranked first if those sentences have important CST relations or if they have a high number of CST relations established among them. The authors concluded that using CST relations improves the quality of the summary.

Afantenos et al. (2004), based on CST, proposes a new classification of relations across documents. The authors divide the relations in two categories: synchronic and diachronic relations. Synchronic relations explore an event being told by different information sources. Diachronic relations, on the other side, explore events that evolve in time for the same information source. Using these new relation classes, the authors propose a summarization methodology that first extracts message templates from the texts (using information extraction tools) and, according to the types of relation that hold among them, produces an unified message that would represent the summary content.

Otterbacher et al. (2002) investigate how CST relations improve cohesion in MDS. They propose the selection of sentences according to content relevance and assume that sentences that have CST relations among them should appear close to each other in the final summary and should be reordered according to possible temporal constraints indicated by CST relations.

In this paper we focus on Radev (2000) MDS proposed method. We formalize, implement and evaluate the MDS redundancy treatment operator and the general operator, as will be detailed in the next section.

3. Methodology

Summarization methodology is traditionally divided in 3 stages: analysis, transformation, and synthesis (Mani and Maybury, 1999). The analysis stage corresponds to the texts understanding, producing an internal representation of their content. The transformation stage performs summarization operations on the internal representation, producing the summary internal representation. In the synthesis stage, the summary internal representation is linguistically realized into the final summary.

In this work, we assume the analysis stage as the CST structuring of the input texts, i.e., annotating CST relations among texts and structuring those relations in a graph. The annotation level is assumed to be the sentence level. Each node of the CST graph represents one sentence and the connections represent the CST relations established among those sentences. We skip this stage by using an already annotated corpus, as will be described latter. The transformation stage corresponds to the content selection task that we are proposing here. From the original CST graph, we select the sentences to be in the summary. Finally, in the synthesis stage, the sentences selected in the previous stage are simply juxtaposed to form the final summary. No rewriting operations are performed. Therefore, we only produce extracts, i.e., summaries built by entire frozen text fragments (sentences, in this case). This is the most common approach to text summarization nowadays.

We explore two methods for content selection in this work. The first method consists in extracting nodes (sentences) that have more CST connections with other sentences, assuming that nodes with more connections are more likely to contain important information. The second method deals with redundancy. Sentences that have some redundant elements will not be

selected to compose the final summary. For example, if two sentences are connected by an equivalence CST relation (which expresses that the two sentences have the same information), only one of the sentences (the shorter one) will be selected for the summary.

We formalize and represent the two content selection methods as content selection operators, following Radev (2000) ideas. The two operators are shown in Figures 3 and 4.

Operator description: generic operator
Relations to keep: all
Relations to deal with: none
Steps: 1. select the sentences that present any CST relation with other sentences 2. rank the selected sentences according to the number of CST relations they present 3. select the best ranked sentences according to the compression rate

Figure 3. Generic content selection operator

Operator description: redundancy operator
Relations to keep: all but identity, equivalence, summary, subsumption
Relations to deal with: identity, equivalence, summary, subsumption
Steps: 1. select the sentences that present any CST relation with other sentences 2. rank the selected sentences according to the number of CST relations they present 3. traverse the rank and analyze the sentences connected by the above relations a. for each identity relation, remove one of the connected sentences, does not mattering which one, since both are identical b. for each equivalence relation, remove the longer sentence connected by it c. for each summary relation, remove the longer sentence connected by it, keeping the sentence that present the summarized information d. for each subsumption relation, remove the subsumed sentence connected by it 4. select the best ranked sentences (from the remaining ones) according to the compression rate

Figure 4. Content selection operator for redundancy treatment

Our operators present four fields: the operator description, for documentation purposes; the "relations to keep" field, which indicates the relations that do not need to be dealt with when selecting the content for the summary; the "relations to deal with" field, which specifies the relations that will have to be managed someway for selecting the appropriate content for the summary; and the "steps" field, which informs step by step how to select the content for the summary considering the relations informed in the previous field.

The meaning of the redundancy relations are as follows. Identity implies that the connected sentences have the same content and that they are written in the same way. Equivalence implies that the connected sentences have the same content written in different ways. Summary relation specifies that the content of one sentence is preserved in the other sentence, but is compressed. Subsumption specifies that the content of a sentence subsumes the content of the other sentence.

Note that the final step in both operators is the selection of the best ranked sentences observing the specified compression rate, i.e., the size of the summary in relation to the size of the source texts. In this paper, we always refer to the size of the longest source text (in number of words). For example, a 70% compression rate specifies that the summary must have at most 30% of the number of words of the longest text in the input group of texts.

It is also important to notice that we are not treating the overlap relation in the second operator. Although such relation also represents redundancy, we are still not able to deal with it. This relation specifies that the related sentences have some content in common, but also have some unique (not shared) content. Dealing with this requires rewriting operations, which we are not considering for the moment. Ideally, one should be able to perform sentence fusion/aggregation of the sentences connected by this relation.

Our general MDS framework is able to read such operators from a file and to perform the summarization according to the user specifications (input texts, compression rate, and preferences – which operations to apply). For the moment, we assume that the CST graph is given as input, but we are already working on an approach for automatic CST analysis (for Brazilian Portuguese language, more specifically).

In the next section we introduce the data we used for evaluating our content selection operators and report our experiments and the results that we obtained.

4. Experiments and Results

For running our experiments, we used a corpus with 30 groups of news texts written in Brazilian Portuguese. Each group has from 2 to 3 texts about the same topic and the corresponding multidocument summary manually built (considering the 70% compression rate over the longest text in the group). The texts were all collected from on-line news agencies and contain in average 20 sentences each one.

The texts in the corpus were annotated by 4 computational linguists who were trained in CST annotation. The corpus annotation took more than 3 months and involved the refinement of CST relations, the construction of an annotation manual, and the development of an annotation tool (Aleixo and Pardo, 2008c). These are themes for other papers, but it is important to say that the annotators agreement was satisfactory, as cited in Section 2.

Two types of evaluation are carried out in this work: automatic evaluation and human evaluation. For the automatic evaluation, we used ROUGE (Lin and Hovy, 2003), an automatic measure for evaluation of summary informativeness. ROUGE receives as input two summaries: one automatic summary (the one we want to evaluate) and (at least) one human summary. ROUGE compares these summaries by basically computing the number of common n-grams. Results are given in terms of precision, recall and f-measure, resulting in figures between 0 (the worst) and 1 (the best, as good as the human summary). On the other hand, human evaluation was carried out to compute the number of redundant elements present in the automatic summaries.

We also compared the results of our content selection methods with the results produced by GistSumm (GIST SUMMarizer) (Pardo et al., 2003; Pardo, 2005), which, to the best of our knowledge, is the only other system available for MDS of Brazilian Portuguese texts. This system is very simple: it merges all texts in only one file and processes it as if it were a single document. Its single document summarization strategy is also very simple: it selects sentences that contain and share the most frequent words in the texts.

Different compression rates were used in both evaluations: 30%, 50%, and 70% over the longest text in each group of texts. The aim of using different compression rates is to evaluate the behavior of redundancy when summaries have different sizes. It is important to notice, however, that the compression rate of the human summary is only one: 70%. This certainly affects the results of the automatic evaluation, but affect equally the results for all the methods, making the comparison of the results still fair.

The average summary informativeness results (obtained from ROUGE evaluation for the corpus groups) are shown in Table 2. In general, it may be observed that CST methods have better results than GistSumm. The generic operator has a considerable improvement in precision, especially when the compression rate is low. This means that this operator produces automatic summaries in which the majority of information correspond to the information present in human summary, in other words, it concentrates a high degree of information. On the other hand, the redundancy operator does not get a better behavior in terms of precision, especially when compression rate is high. We believe that, for this operator, when summary becomes smaller, it is highly probable that important information is ignored. For methods that do not treat redundancy (as GistSumm), this is less probable to occur, since there are more

redundant information in the selected content and, therefore, some of them will probably be in the summary.

Table 2. Results for the informativeness evaluation

	Compression rate	Precision	Recall	F-measure
Generic operator	30%	0.79000	0.30877	0.43723
Generic operator	50%	0.75768	0.35363	0.45481
Generic operator	70%	0.62287	0.45191	0.53739
Redundancy operator	30%	0.50853	0.55612	0.51126
Redundancy operator	50%	0.43619	0.57525	0.45884
Redundancy operator	70%	0.22086	0.59809	0.31315
GistSumm	30%	0.55621	0.31295	0.38674
GistSumm	50%	0.50312	0.37856	0.41278
GistSumm	70%	0.42651	0.44752	0.43595

Table 2 shows the average number of redundant information for each method and each compression rate.

Table 3. Results for redundancy evaluation

	Compression rate	Number of redundant sentences
Generic operator	30%	9
Generic operator	50%	7
Generic operator	70%	5
Redundancy operator	30%	3
Redundancy operator	50%	1
Redundancy operator	70%	0
GistSumm	30%	10
GistSumm	50%	7
GistSumm	70%	6

As it can be seen in Table 3, redundant sentences in the automatic summaries are significantly higher for those methods which do not treat redundancy in particular. The redundant elements appear especially when compression rate is low and consequently the size of the final summary is bigger. This explains the better results for redundancy operator in terms of recall.

5. Conclusions and Final Remarks

In this paper we explored a knowledge-based approach for content selection. Differently from works that use superficial measures for selecting the relevant sentences to put in the summary, we used CST relations. CST relations help to explore properties among various texts in terms of similarities and differences. Knowing these properties allows a better treatment of texts. In this paper, we demonstrate it for multidocument summarization. Particularly, CST-based methods show an improvement in dealing with factors such as redundancy and information quality. Compression rate is also an important factor in summary quality, since it cannot be too high or too low, which makes the employment of CST information less useful. If the summary is too small, there is probably space for only the main sentence, being not necessary to have CST relations for including other sentences; on the other side, if the summary is too big, most of information will already be in the summary and CST will be only useful for dealing with some specific types of phenomena, for instance, redundancy.

Future work shall explore other content selection operators, including their combination with non CST-based strategies, producing hybrid approaches. We also plan to use some sentence fusion tool for dealing with some CST relations, for instance, the one presented by Seno and Nunes (2009). In the future, when a complete CST parser (under construction) is available, the CST-based MDS process may be completely automatic.

Acknowledgments

The authors are grateful to FAPESP, CNPq, and CAPES for supporting this work.

References

Afantenos, S.D.; Doura, I.; Kapellou, E.; Karkaletsis, V. (2004). Exploiting Cross-Document Relations for Multi-document Evolving Summarization. In the *Proceedings of SETN*, pp. 410-419.

Aleixo, P. and Pardo, T.A.S. (2008a). Finding Related Sentences in Multiple Documents for Multidocument Discourse Parsing of Brazilian Portuguese Texts. In *Anais do VI Workshop em Tecnologia da Informação e da Linguagem Humana* – TIL, pp. 298-303. Vila Velha, Espírito Santo. October, 26-28.

Aleixo, P. and Pardo, T.A.S. (2008b). *CSTNews: Um Córpus de Textos Jornalísticos Anotados segundo a Teoria Discursiva Multidocumento CST (Cross-document Structure Theory)*. Série de Relatórios Técnicos do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, N. 326.

Aleixo, P. and Pardo, T.A.S. (2008c). *CSTTool: Uma Ferramenta Semi-automática para Anotação de Córpus pela Teoria Discursiva Multidocumento CST*. Série de Relatórios Técnicos do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, N. 321.

Lin, C.Y. and Hovy, E. (2003). Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics. In the *Proceedings of 2003 Language Technology Conference*. Edmonton, Canada.

Mani, I. (2001). *Automatic Summarization*. John Benjamins Publishing Co. Amsterdam.

Mani, I. and Maybury, M. T. (1999). *Advances in automatic text summarization*. MIT Press, Cambridge, MA.

Mann, W.C. and Thompson, S.A. (1987). *Rhetorical Structure Theory: A Theory of Text Organization*. Technical Report ISI/RS-87-190.

Otterbacher, J.C.; Radev, D.R.; Luo, A. (2002). Revisions that improve cohesion in multi-document summaries: a preliminary study. In the *Proceedings of the Workshop on Automatic Summarization*, pp 27-36.

Pardo, T.A.S.; Rino, L.H.M.; Nunes, M.G.V. (2003). GistSumm: A Summarization Tool Based on a New Extractive Method. In the *Proceedings of the 6th Workshop on Computational Processing of the Portuguese Language - Written and Spoken* – PROPOR (Lecture Notes in Artificial Intelligence 2721), pp. 210-218. Faro, Portugal.

Pardo, T.A.S. (2005). *GistSumm - GIST SUMMarizer: Extensões e Novas Funcionalidades*. Série de Relatórios do NILC. NILC-TR-05-05. São Carlos-SP/Brasil.

Radev, D.R. and McKeown, K. (1998). Generating natural language summaries from multiple on-line sources. *Computational Linguistics*, Vol. 24, N. 3, pp. 469-500.

Radev, D.R. (2000). A common theory of information fusion from multiple text sources, step one: Cross-document structure. In the *Proceedings of the 1st ACL SIGDIAL Workshop on Discourse and Dialogue*.

Seno, E.R.M. and Nunes, M.G.V. (2009). Reconhecimento de Informações Comuns para a Fusão de Sentenças Comparáveis do Português. *Linguamática*, Vol. 1, pp. 71-87.

Trigg, R. (1983). *A Network-Based Approach to Text Handling for the Online Scientific Community*. Ph.D. Thesis. Department of Computer Science, University of Maryland.

Trigg, R. and Weiser, M. (1987). TEXTNET: A network-based approach to text handling. *ACM Transactions on Office Information Systems*, Vol. 4, N. 1, pp. 1-23.

Zhang, Z.; Goldenshon, S.B.; Radev, D.R. (2002). Towards CST-Enhanced Sumarization. In the *Proceedings of the 18th National Conference on Artificial Intelligence*.