Proceedings of Symposium in Information and Human Language Technology. Natal, RN, Brazil, November 4–7, 2015.

On Strategies of Human Multi-Document Summarization

Renata T. Camargo[1], Ariani Di-Felippo[1], Thiago A. S. Pardo[2]

Núcleo Interinstitucional de Linguística Computacional (NILC)

[1]Departamento de Letras – Universidade Federal de São Carlos
Caixa Postal 676 – 13565-905 – São Carlos – SP – Brazil

[2]Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo
Caixa Postal 668 – 13566-970 – São Carlos – SP – Brazil

renatatironi@hotmail.com, arianidf@gmail.com, taspardo@icmc.usp.br

Abstract. *In this paper, using a corpus with manual alignments of human-written summaries and their source news, we show that such summaries consist of information that has specific linguistic features, revealing human content selection strategies, and that these strategies produce indicative results that are competitive with a state of the art system for Portuguese.*

Resumo. *Neste artigo, a partir de um corpus com alinhamentos manuais entre sumários e suas respectivas notícias-fonte, evidencia-se que tais sumários são compostos por informações que possuem características linguísticas específicas, revelando estratégias humanas de sumarização, e que essas estratégias produzem resultados iniciais que são competitivos com um sistema do estado da arte para o português.*

1. Introduction

The increasing of new technologies has had an impact on the amount of available textual information on the web. Consequently, Multi-document Summarization (MDS) appears to be a useful Natural Language Processing (NLP) application to promote quick access to large quantities of information, since it produces a unique summary from a collection or cluster of texts on the same topic or related topics [Mani 2001]. Within a generic perspective, the multi-document summary should ideally contain the most relevant information of the topic that is being discussed in the source texts. Moreover, MDS should not only focus on the extraction of relevant information, but also deal with the multi-document challenges, such as redundant, complementary and contradictory information, different writing styles and varied referential expressions.

There are two ways of approaching MDS [Mani 2001]. The superficial/shallow approach uses little linguistic information (or statistics) to build summaries. The deep approach is characterized by the usage of deep linguistic knowledge, i.e., syntactic, semantic or discourse information. The superficial approach usually requires low-cost processing, but produces summaries that tend to have lower linguistic quality. The deep approach is said to produce summaries of higher quality in terms of information, coherence and cohesion, but it demands various high-cost resources. The deep and superficial MDS applications commonly produce extracts (i.e., summaries generated by concatenating sentences taken exactly as they appear in the source texts), but deep approach can also generate abstracts (i.e., with rewriting operations).

To select the sentences to compose the summaries, MDS may take into account human strategies from single-document summarization, codified in features such as sentence position and word frequency [Kumar and Salim 2012]. Regarding human multi-document summarization (HMDS), only redundancy has been widely applied as criterion for content selection, which is based on the empirical observation that the most repeated information covers the main topic of the cluster [Mani 2001; Nenkova 2006].

In this context, this work is focused on the investigation of HMDS content selection strategies. Particularly, for a corpus of news texts, we study some superficial and deep sentence features that may be useful for summarization. Since the source sentences in this corpus are aligned to the sentences of the correspondent reference (human) summary, we show that a machine learning technique could identify that a few features characterize well the aligned sentences (i.e., the sentences whose content was selected to the summary), achieving 70.8% of accuracy. We also show that additional experiments with the best learned HMDS strategy indicated that it may produce competitive results with a state of the art system for Portuguese, outperforming it for a small test corpus. Consequently, this work contributes to the understanding of the HMDS task and to the improvement of the automatic process by providing linguistic insights.

To describe this work, we organized the paper in 5 sections. In Section 2, we describe the main human content selection strategies and the correspondent features of the literature. In Section 3, the used methodology is reported. In Section 4, results are discussed, and, in Section 5, some final remarks are made.

2. Human Content Selection in Text Summarization

In one of the most comprehensive study of human summarization, Endres-Niggemeyer (1998) established that humans perform single-document summarization in three stages: (i) document exploration, (ii) relevance assessment, and (iii) summary production. This means that humans first interpret the source-text, then select important information from it, and finally present a new text in the form of a summary.

Regarding the relevance assessment stage, where, according to Hasler (2007), humans perform the core summarization task (i.e., the selection of the relevant information), Endres-Niggemeyer pointed out the use of some strategies. Some well-known shallow features are [Kumar and Salim 2012]:

(i) *sentence length* or *size*, according to which very short or long sentences may not be suitable to compose the summary;
(ii) *sentence position*, according to which sentences in the initial positions of a text should compose a summary;
(iii) *word frequency*, according to which the summary is produced by retrieving and putting together sentences with the highest frequent content words in the cluster;
(iv) *title/subtitle word*, according to which the relevance of a sentence is the sum of all the content words appearing in the title and (sub-)headings of their text, and;
(v) *cue word/expression*: according to which the relevance of a sentence is computed by the presence or absence of certain cue words or expressions.

Although multi-document summarization can be conceived as an extension of the single one, humans seem to use specific strategies for relevance assessment in the scenario of multiple source texts, which have been empirically observed and reported in MDS

literature. The main one is the selection of the most redundant information in a collection to produce the corresponding summary, as we have already mentioned before [Mani 2001; Nenkova 2006]. The other is that humans choose one text of their preference as a basis to select the main information and then they seek the other texts of the cluster to complement the multi-document summary information [Mani 2001; Camargo 2013]. For the choice of the basis source text, many linguistic or extra-linguistic factors may influence, such as: (i) date of publication (i.e., humans can first consider the latest or the oldest text, depending on the interest), (ii) prestige of the journalistic vehicle, etc.

In feature-based methods of MDS, word frequency may indicate redundancy. In other shallow methods, such as those based on clustering, highly similar sentences of a collection are grouped into one cluster, which generates a number of clusters. A very populous cluster represents redundant information or topic. Hence, for each of the most populous clusters, the methods select only one sentence to compose the summary, which is based on the closeness of the sentence to the centroids (i.e., frequent occurring words) of the cluster. In graph-based methods, the source documents are represented in a graph where each sentence becomes a node and the weighted connections between nodes codify the similarity between the corresponding sentences. A redundant sentence is the one that is strongly connected to other sentences.

In deep approaches, semantic-based MDS methods commonly map nouns of the input sentences onto concepts of a hierarchy or ontology, and then select the sentences with the most frequent concepts of the collection to produce the summary (e.g. Lin et al. (2010)). Discourse-based methods take into account discourse relations such as those of the *Cross-document Structure Theory* (CST) [Radev 2000]. These works represent the input texts in a graph, where each node codifies one sentence and the connections represent the CST relations established among those sentences. For content selection, one method consists in extracting sentences that have more CST connections with other sentences, assuming that they are redundant and, then, more relevant.

In this paper, we test features from the above approaches to look for a good summarization strategy. We describe the method used in this work in the next section.

3. Corpus-based Investigation of HMDS strategies

The experiments in this work were conducted over CSTNews corpus [Cardoso et al. 2011], a multi-document corpus that is composed of 50 clusters of news texts in Brazilian Portuguese. Each cluster contains 2 or 3 news texts on the same topic, automatic and human multi-document summaries (with a 70% compression rate[1]), and many annotation layers. In this corpus, each sentence of the input texts is aligned to one or more sentences of the correspondent human multi-document summary, which indicates the origin of the summary content. The manual alignment was performed in the summary-to-text direction according to content overlap rules [Camargo et al. 2013; Agostini et al. 2014]. To illustrate, the summary sentence (1), "*17 people died after a plane crash in the Democratic Republic of Congo*", is aligned to the text sentence (2), "*A crash in the town of Bukavu in the eastern Democratic Republic of Congo (DRC), killed 17 people on Thursday afternoon, said on Friday a spokesman of the United*

[1] This rate means that the summary may have up to 30% of the number of words of the longest text of the cluster.

Nations". Approximately 78% of the summary sentences were aligned to more than one sentence of the source texts.

Having this corpus, our investigation followed the following stages: feature selection, corpus description (in terms of the features), and HMDS strategy identification. From the literature, we used 8 features as strong indicators for content selection in HMDS: 4 shallow and 4 deep features.

The shallow features correspond to characteristics that refer to the structure of the text or sentence. Particularly, we selected 4 features: *size, frequency, keyword,* and *position*[2]. In our experiments, the values of the first 3 features are normalized in order to avoid discrepancies in the data due to cluster variations. We use the previous parsing annotation of CSTNews, generated by PALAVRAS [Bick 2000], for computing size, frequency, and keyword features.

The sentence size describes the size or length of a sentence in terms of the number of content words it contains. The normalized size is the ratio of the number of words occurring in the sentence over the number of words occurring in the longest sentence of the cluster. For example, the sentence 6 from document 1 of the cluster 9 (S6D1C9), "*The others will be in Rondônia"*, has 2 content words, "be" and "Rondônia". Considering that the longest sentence in cluster 9 is composed by 43 content words, the normalized size of S6D1C9 is 2/43=0.046.

The frequency of a sentence is the sum of the frequency (in the cluster) of the content words it contains. To normalize the feature, we divide the value of a sentence by the highest frequency value of a sentence in the cluster. For example, the frequency value of S6D1C9 is 18 because the frequencies of "be" and "Rondônia" in cluster 9 are, respectively, 1 and 17. Given that the highest frequency obtained by a sentence in the same cluster is 230, the normalized frequency of S6D1C9 is 18/230=0.078.

The keyword feature of a sentence is computed as the sum of the 10% most frequent content words in the cluster that occur in the sentence. To normalize the feature, we divide the keyword value of each sentence by the highest keyword value of the cluster. For instance, the S6D1C9 has only 1 keyword, "Rondônia". Thus, the normalized keyword value of S6D1C9 is 0.05 because 1 is divided by 20, which is the highest keyword value in the cluster. It is worth noting that frequency and keywords are different superficial techniques that may indicate redundancy.

The sentence position refers to the location of the sentence in the source text. This feature can assume 3 possible values: *begin*, *middle*, and *end*. *Begin* value corresponds to the first sentence of the text, *end* value corresponds to the last sentence, and *middle* corresponds to the remaining sentences between "begin" and "end".

The deep feature set refers to discourse characteristics of the texts provided by the annotation of the corpus with CST (*Cross-document Structure Theory*) [Radev 2000]. For the manual annotation, 14 CST relations were used (namely, *Identity, Equivalence, Summary, Subsumption, Overlap, Follow-up, Historical background*, *Elaboration, Contradiction, Citation, Attribution, Modality, Indirect speech, and Translation*). Considering sentences as the basic segments, we illustrate an *Equivalence*

[2] We did not consider the other popular features of the literature, as *title word* and *cue word,* because CSTNews do not provide the title for all source texts and such cue words are more suitable for scientific texts.

(paraphrasing) with the following two sentences from different texts [Radev 2000, p. 79]: "*Ford's program will be launched in the United States in April and globally within 12 months"* and "*Ford plans to introduce the program first for its employees in the United States, then expand it for workers abroad*". The CST annotation of a cluster in CSTNews is a graph, whose nodes are sentences and the edges are relations. The nodes may be disconnected, since not all sentences present relations with others.

According to the CST typology proposed by Maziero et al. (2010), we specified 4 features: *redundancy, complement, contradiction* and *form*. The redundancy feature of a sentence corresponds to the number of the following CST relations that the sentence presents: *Identity, Equivalence, Overlap, Summary,* and *Subsumption*. The complement feature corresponds to the number of *Historical background, Elaboration* and *Follow-up* relations. The contradiction feature is the number of *Contradiction* relations. Finally, the form feature codifies the number of *Citation, Attribution, Modality, Indirect-speech* and *Translation* relations. To normalize these features in a specific cluster, we divide the feature value by the total number of relations in the cluster. As an example of how these features are calculated, consider a sentence that is connected by a *Subsumption* and an *Attribute* relation to other sentences. This sentence has 1 relation of the redundancy category and 1 of the form category. Supposing that these are the only relations in the cluster, the sentence has the following feature-value pairs: redundancy=0.5 (=1/2), complement=0, contraction=0, and form=0.5 (=1/2).

Once the features for each sentence in the source texts were computed, we need to determine the correspondent class of each sentence in our corpus. Since for each cluster in CSTNews we have the summary-text alignments, we can determine which sentences had their content selected for the summary. Two possible classes can be assigned: "yes" or "no". Sentences classified as "yes" represent the ones that were aligned to the summary and sentences classified as "no" represent the ones that were not aligned (and, therefore, were considered irrelevant to be included in the summary).

To perform the machine learning over CSTNews, we applied the *10-fold cross validation*[3] technique, which gets more realistic estimates of the error rates for classification, since our dataset is relatively small. In total, there are 2080 learning instances in our dataset, with 57% of them belonging to the "no" class, which, in summarization, is usually the majority class. We used Weka environment [Witten and Frank 2005] for running all the algorithms, and general accuracy for evaluating the results.

Our focus in this paper is to look for symbolic approaches to the task, given that, more than a good classification accuracy, we want to be able to make the summarization strategy explicit. Nonetheless, we have also tested other machine learning techniques from other paradigms, for comparison purposes only. We explore in more details the results achieved by the symbolic approaches, and only briefly comment on the results of the other approaches that we consider, i.e., the connectionist and mathematical/probabilistic approaches.

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

In the connectionist paradigm, we used the well known method called Multi-Layer Perceptron (MLP), with the default Weka configurations. We achieved 65.7% of accuracy. Among the several mathematical/probabilistic methods in Weka, we run Naïve-Bayes and SMO. Naive-bayes achieved 69% of accuracy, while SMO was the highest among all the algorithms, achieving 70.9%.

The symbolic methods produce rules/trees that can be verified by human experts. Among them, we tried JRip, PART, Prism, J48, and OneR. PART and Prism algorithms generated long sets of rules (more than 60) with close accuracy (approximately 69%). The decision tree produced by J48 also contains many rules, but it presents slightly higher accuracy, 70.2%. OneR algorithm uses only the most discriminative feature to produce a unique set of rules over this feature. In our case, OneR selected the redundancy feature and achieved 70.5% of accuracy. As usual, it surprisingly produced very good results, but did not outperform JRip, which we discuss below.

JRip learned a small set of rules with the best accuracy, 70.8%. Such combination (manageable rule set and highest accuracy among the symbolic approaches) makes the choice of JRip a good one for our purposes. Table 1 presents the 9 rules of JRip, which are followed by the number of instances (sentences) correctly classified and incorrectly classified, and the precision of the rule, given by the number of correctly classified instances over all the instances classified by that rule.

Table 1 – JRip logic rules

Rules	Correct	Incorrect	Precision (%)
1. If Position = beginning then "yes"	140	16	89.7
2. Elseif Redundancy = 0.9-inf then "yes"	81	11	88
3. Elseif Redundancy = 0.3-0.5 then "yes"	369	164	69.2
4. Elseif Redundancy = 0.6-0.8 then "yes"	114	19	85.7
5. Elseif Redundancy = 0.2-0.3 and Frequency = 0.5-0.6 then "yes"	35	9	79.5
6. Elseif Redundancy = 0.1-0.2 and Frequency = 0.4-0.5 then "yes"	10	2	83.3
7. Elseif Redundancy = 0.1-0.2 and Size = 0.2-0.3 then "yes"	12	2	85.7
8. Elseif Size = 0.1-0.2 and Frequency = 0.3-0.4 then "yes"	14	3	82.3
9. Elseif "no"	1305	346	79

In the rules, one can say that position, redundancy, frequency and size features characterize well the aligned sentences of CSTNews, i.e., sentences whose content composes the summary. As it is well known about position, the "beginning" value in Rule 1 reveals that human commonly select the first sentences of source documents to compose a summary. We may justify this strategy by the "inverted pyramid" structure of news, in which the first sentence conveys the primary information ("lead"). Redundancy (codified by CST relations or word frequency) is the most characteristic feature, since 7 of the 9 rules are based on it, individually or in combination with other features. For attribute selection[4] was applied two methods (at Weka), i.e., InfoGainAttributeEval and CfsSubsetEval, and both indicated the relevance of the

[4] The aim of attribute selection is to improve the performance of the algorithms. It is important because there are attributes that can be irrelevant and removing them can reduce the processing time and generate simpler models.

redundancy feature. Thus, selecting the most repeated information as a HMDS strategy is confirmed in our corpus investigation. Moreover, the low values of the size feature indicate that humans select content preferably expressed by medium or short sentences. The above results demonstrate that the human single-document summarization strategies based on position, frequency and size are also applied in HMDS. If none of the 8 first rules are applied, the default class is "no" (i.e., non-aligned sentence), which is given the 9th rule.

It is also interesting to see how productive the rules are. For instance, rules 1 to 4 deal with many more cases than rules 6 to 8, which is natural to happen due to the way the machine learning process chooses the features to start the rules. Given that, one might still achieve good results by using only the first 4 rules for the "yes" class and the last default rule for the "no" class. In Table 2, we have the JRip confusion matrix, by means of which we verify in more details how the classifier is dealing with each class. Each column of the matrix represents the instances in a predicted class, while each row represents the instances in an actual class.

Table 2 - Confusion matrix of JRip algorithm

Class \ Test	Aligned (895) "Summary=yes"	Non-aligned (1185) "Summary=no"
Aligned	523	372
Non-aligned	235	950

It can be observed from the results of Table 2 that, from the total of 1185 non-aligned source sentences, the rules of JRip correctly classified 950 of them. Still, from the total of 895 sentences of source texts that were aligned to the summaries, the algorithm correctly identified 523 of them. Based on this performance, we may conclude that JRip correctly classified more non-aligned sentences than aligned ones. This might be a consequence of the aforementioned unbalanced nature of our training corpus. It is important to say that we opted for not balancing the data (by using oversampling, for instance), since the task is naturally unbalanced in the real world.

We now describe the evaluation of summaries produced by the JRip rules.

4. Summary Evaluation

Besides the results of the machine learning, we were also interested in checking the quality and informativeness of the summaries produced by the JRip rules. These are the criteria that are usually assessed in summaries.

In order to do this summary evaluation, we manually created a new test corpus with the same characteristcs of CSTNews. The test corpus consists of 6 clusters, and each of them contains: (i) 3 news texts on the same topic, (ii) 3 human multi-document summaries (abstracts), produced by different computational linguists, with 70% compression rate, (iii) sentential alignments among source texts and human summaries, and (iv) CST annotation in the texts. We restricted the corpus to only 6 clusters because text annotation and summary writing tasks are expensive and time consuming tasks.

The summary building process is as follows. Given a cluster of the test corpus, we first apply the JRip rules to select the sentences that are worthy to be in the summary (only sentences classified as "yes" are considered). Having these "yes" sentences, we

need to rank them in order to produce a sentence relevance rank, which we do by ordering the sentences by the precision of the rule that was applied to select each sentence (see Table 1); if it happens that there are sentences competing for the same position in the rank (supposing that the rules that selected them had the same precision), we give preference to sentences that come first in their texts; if this is not enough to distinguish them (supposing that they are in the same position in different texts), we order them by the prestige of the source, as indicated by Camargo (2013). Having the rank, we start selecting the best ranked sentences to compose the summary, always checking for redundancy between the newly selected sentence and eventual previously selected sentences to the summary. We use the information provided by CST to eliminate redundancy, by discarding the candidate sentence that has relations of the redundancy category with the ones already selected to the summary. For example, if the relation between two sentences is *Identity*, the new sentence is ignored; if the relation is *Equivalence*, we eliminate the longest sentence (considering the number of words in the sentence); if the relation is *Subsumption*, we eliminate the sentence that is subsumed. We select as many sentences to the summary as the compression rate allows.

To analyze the quality of the summaries, we used the 5 traditional criteria proposed by the DUC conference [Dang 2005]: (i) gramaticality (G): the summary should have no datelines, capitalization errors or ungrammatical sentences; (ii) non-redundancy (NR): there should be no unnecessary repetition in the summary; (iii) referential clarity (RC): it should be easy to identify who or what the pronouns and noun phrases in the summary are referring to; (iv) focus (F): the summary should only contain information that is related to the rest of the summary, and (v) structure and coherence (SC): the summary should be well-structured and well-organized, i.e., it should not just be a heap of related information.

For comparison, the summaries generated by another method of MDS for the same 6 clusters were also judged considering the same textual properties. In this case, the automatic method used to generate the comparison summaries was RSumm [Ribaldo et al. 2012], which is one of the state of the art systems for Portuguese.

The evaluation of the properties related to quality was performed by 10 computational linguists. For each automatic summary, the judges scored each of the 5 textual properties through an online form. For all properties, judges had a scale from 1 to 5 points, being 1=very poor, 2=poor, 3=barely acceptable, 4=good, and 5=very good. The results are shown in Table 3. The values are presented in two ways: (i) absolute values (which is the number of votes for the corresponding scale), and (ii) percentage. Looking to the average values, one may see that the JRip rules outperform Rsumm in all the evaluated criteria, indicating that the used features in this study are better at dealing with textuality factors in the summaries.

Regarding informativeness evaluation, we used the traditional automatic ROUGE (Recall-Oriented Understudy for Gisting Evaluation) measure [Lin 2004], which is mandatory in the area. ROUGE computes the number of common n-grams among the automatic and reference/human summaries, being able to rank automatic summaries as well as humans would do, as its author has shown. Table 4 shows average ROUGE results for 1-grams (referenced by ROUGE-1), 2-grams (ROUGE-2) and the longest common subsequence (ROUGE-L) overlap, in terms of Recall (R), Precision (P) and F-measure (F), for both JRip rules and RSumm. Basically, recall computes the

amount of common n-grams in relation to the number of n-grams in the reference summaries; precision computes the number of common n-grams in relation to the n-grams in the automatic summary; the f-measure is the harmonic mean of the previous 2 measures, being an unique indicator of the system performance. One may see that the JRip rules outperform RSumm in all the measures. If we consider the f-measure for ROUGE-1, which is by far the most used in the literature, we may see that the JRip rules are approximately 6.7% better than RSumm.

Table 3. Linguistic quality evaluation of summaries with DUC criteria

Criteria	Method	Very poor (1)		Poor (2)		Barely Acceptable (3)		Good (4)		Very Good (5)		**Average**
G	HMDS	0	**0%**	0	**0%**	3	**5%**	18	**30%**	39	**65%**	4,7 (**very good**)
	RSumm	0	**0%**	0	**0%**	7	**11,6%**	22	**36,6%**	31	**51,6%**	4,4 (**good**)
NR	HMDS	0	**0%**	0	**0%**	2	**3,3%**	15	**25%**	43	**71,6%**	4,7 (**very good**)
	RSumm	0	**0%**	2	**3,3%**	17	**28,3%**	17	**28,3%**	24	**40%**	4,1 (**good**)
RC	HMDS	0	**0%**	0	**0%**	9	**15%**	20	**33,3%**	31	**51,6%**	4,4 (**good**)
	RSumm	0	**0%**	2	**3,3%**	5	**8,3%**	26	**43,3%**	27	**45%**	4,3 (**good**)
F	HMDS	0	**0%**	0	**0%**	3	**5%**	24	**40%**	33	**55%**	4,5 (**very good**)
	RSumm	1	**1,6%**	4	**6,6%**	11	**18,3%**	22	**36,6%**	22	**36,6%**	4,0 (**good**)
SC	HMDS	0	**0%**	0	**0%**	7	**11,6%**	33	**55%**	20	**33,3%**	4,2 (**good**)
	RSumm	0	**0%**	6	**10%**	19	**31,6%**	23	**38,3%**	12	**20%**	3,7 (**good**)

Table 4. Informativeness evaluation of summaries with ROUGE

	Avg. ROUGE-1			**Avg. ROUGE-2**			**Avg. ROUGE-L**		
	R	**P**	**F**	**R**	**P**	**F**	**R**	**P**	**F**
JRip rules	0.444	0.517	0.464	0.200	0.242	0.212	0.373	0.441	0.392
RSumm	0.425	0.448	0.435	0.191	0.202	0.196	0.358	0.378	0.367

It is important to say, however, that such results are only indicative of what we may expect from the rules and the discriminative power of the studied features, since the test set for quality evaluation and ROUGE was too small (only 6 clusters). For a more reliable result, we would need to run the rules for a bigger corpus. We could not do that for CSTNews because this corpus was already used for creating the rules (during the training), and using it for testing would result in a biased evaluation. And, besides CSTNews, we are not aware of other corpora with the data/annotation we need for our rules to work.

Having the reservations been made, it is interesting that the rules could outperform RSumm (even for a small test corpus), since highly deeper and more informed approaches have struggled to do that (see, e.g., Cardoso (2014)). This shows how effective the learned HDMS strategy is.

5. Final Remarks

To the best of our knowledge, this integrated study of features over a corpus of human summaries and their application in an automatic method is new in the area and, at least for Portuguese, has potential to advance the known state of the art. Future work may include the study of other features, as well as a more detailed characterization of the summaries, in terms of lexical and syntactical patterns.

References

Agostini, V.; Camargo, R.T.; Di-Felippo, A.; Pardo, T.A.S. (2014). Manual alignment of news texts and their multi-document human summaries. In Aluísio, S.M. and Tagnin, S.E.O. (Eds.), *New language technologies and linguistic research: a two-way road*, pp. 148-170. Cambridge: Cambridge Scholars Publishing.

Bick, E. (2000). *The Parsing System Palavras - Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Famework*. PhD Thesis. Aarhus University Press.

Camargo, R.T.; Agostini, V.; Di-Felippo, A.; Pardo, T.A.S. (2013). Manual typification of source texts and multi-document summaries alignments. *Procedia - Social and Behavioral Sciences*, Vol. 95, pp. 498-506.

Camargo, R.T. (2013). *Investigação de Estratégias de Sumarização Humana Multidocumento*. Dissertação de Mestrado. Universidade Federal de São Carlos. 135p.

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. (2011). CSTNews - A Discourse-Annotated Corpus for Single and Multi-Document Summarization of News Texts in Brazilian Portuguese. In the *Proceedings of the 3rd RST Brazilian Meeting*, pp. 88-105.

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

Dang, H.T. (2005). Overview of DUC 2005. In the *Proceedings of the Document Understanding Conference*.

Endres-Niggemeyer, B. (1998). *Summarization Information*. Berlin: Springer.

Hasler, L. (2007). From extracts to abstracts: human summary production operations for Computer-Aided Summarisation. In the *Proceedings of the RANLP Workshop on Computer-aided Language Processing*, pp. 11-18.

Kumar, Y.J.; Salim, N. (2012) Automatic Multi-Document Summarization Approaches. *Journal of Computer Science 8* (1): 133-140. ISSN 1549-3636

Li, L., D. Wang, C. Shen; T. Li (2010). Ontology enriched multi-document summarization in disaster management. *Proceedings of the 33rd international ACM SIGIR*, July 19-23, ACM, New York, USA, pp. 820. ISBN: 978-1-4503-0153-4

Lin, C-Y. (2004). ROUGE: a Package for Automatic Evaluation of Summaries. In the *Proceedings of the Workshop on Text Summarization Branches Out*.

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

Maziero, E.G.; Jorge, M.L.C.; Pardo, T.A.S. (2010). Identifying Multidocument Relations. In the *Proceedings of the 7th International Workshop on Natural Language Processing and Cognitive Science*, pp. 60-69.

Nenkova, A. (2006). *Understanding the process of multi-document summarization: content selection, rewrite and evaluation*. PhD Thesis. Columbia University.

Radev, D. R. (2000). A common theory of information fusion from multiple text sources, step one: cross-document structure. In the *Proceedings of the ACL SIGDIAL Workshop on Discourse and Dialogue*, pp. 74-86.

Ribaldo, R.; Akabane, A.T.; Rino, L.H.M.; Pardo, T.A.S. (2012). Graph-based Methods for Multi-document Summarization: Exploring Relationship Maps, Complex Networks and Discourse Information. In the *Proceedings of the 10th International Conference on Computational Processing of Portuguese* (LNAI 7243), pp. 260-271.

Witten, I.H. and Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.