2017 Brazilian Conference on Intelligent Systems

Update Summarization for Portuguese

Fernando Antônio Asevedo Nóbrega and Thiago Alexandre Salgueiro Pardo
Interinstitutional Center for Computational Linguistics (NILC)
Institute of Mathematical and Computer Sciences, University of São Paulo
{fasevedo,taspardo}@icmc.usp.br

***Abstract*—Update summarization aims at automatically producing a summary from a collection of texts for a reader that already has some previous knowledge about the subject. It is a challenging task, since it not only brings all the demands from the summarization area (as producing informative, coherent and cohesive summaries) but also includes the issue of reporting only relevant new/updated content. In this paper, we report a comprehensive investigation of update summarization methods for the Portuguese language, for which there are few initiatives, and propose a new method that combines the summarization strategies of some other methods, producing better results and advancing the state of the art. More than this, we introduce a reference dataset and establish an experiment setup in the area in order to foster future research.**

***Index Terms*—Summarization, empirical and statistical methods, evaluation**

I. Introduction

The Update Summarization (US) task aims at producing a summary from a collection of related (source) texts/documents under the assumption that the reader has some previous knowledge about the subject of the texts. This happens, for instance, when the reader has previously read some related material. Thus, it is expected that an update summary be produced with just the most relevant new or more recent information in order to update the user about the subject.

The US task is a natural evolution of the traditional Automatic Summarization tasks [1], as the Multi-Document Summarization [2], which aims to better handle the current online environment, in which a huge amount of data and new content is very quickly produced and in many different sources, producing a situation in that the knowledge that a reader has about a specific topic may be easily outdated.

Figures 1 and 2 show examples of a regular (generic) summary and an update summary with no more than 100 words, respectively (in Portuguese, which is the original language). The same text collection was used as source for both of them. In the case of the update summary, it was assumed that the reader has already read another text collection on the same subject. The source texts are about a past update in the National Agency of Civil Aviation in Brazil (Anac).

One may see that they have different shapes as they serve to different purposes. It is interesting to notice the several challenges that the US task brings. It includes the challenges that come from the traditional summarization area, as producing informative summaries that are coherent and cohesive, dealing with the multi-document phenomena (as the occurrence of redundant, contradictory and complementary information in the texts) and temporally ordering the events and facts, among many others. Furthermore, the US task introduces new challenges, as modeling the user previous knowledge (which is generally made by considering that the user knowledge is unstructured and represented in a collection of texts that was previously read) and finding relevant new information to compose a summary.

O ministro da Defesa, Nelson Jobim, deve encaminhar o nome da economista Solange Vieira para assumir uma das diretorias da Agência Nacional de Aviação Civil (Anac). Ainda não está definida a diretoria que a economista vai assumir. O ministro da Defesa, Nelson Jobim, informou no fim da noite desta terça-feira que a economista Solange Vieira, de 38 anos, será a nova presidente da Agência Nacional de Aviação Civil (Anac). - A Solange vai ser a nova presidente da Anac - disse Jobim, em jantar que celebrou os 50 anos da Rede RBS em Brasília. Mas, diante da dificuldade para encontrar

Fig. 1. An example of a regular (generic) summary.

O ministro da Defesa, Nelson Jobim, informou no fim da noite desta terça-feira que a economista Solange Vieira, de 38 anos, será a nova presidente da Agência Nacional de Aviação Civil (Anac). O relatório final da CPI do Apagão da Câmara, que começou a ser lido nesta terça-feira, recomenda o ingresso da iniciativa privada na administração da infra-estrutura dos aeroportos, hoje sob o comando de uma estatal, a Infraero. Ele disse que não está convencido da ”participação objetiva” de Zuanazzi nas denúncias contra a agência: - Não podemos indiciar para agradar à oposição, ao governo ou a quem quer que seja

Fig. 2. An example of an update summary.

The area has a relatively short history. The US task was formally introduced in a pilot track at the Document Understand Conference (DUC[1]) in 2007. US has also been present in the Text Analysis Conferences (TAC[2], in a new incarnation of DUC conferences) since 2008. In DUC 2007, each test set had 3 text collections, named A, B and C, which were sorted by their respective timestamps. A summary with no more than 100 tokens (whitespace tokenized) should be produced for each one of them, considering that, for a collection i, it is assumed that the reader knew the previous ones [3]. For instance, it was assumed that the reader had already

[1] at http://duc.nist.gov/duc2007/tasks.html
[2] at http://tac.nist.gov

978-1-5386-2407-4/17 $31.00 © 2017 IEEE
DOI 10.1109/BRACIS.2017.49

read the A and B collections when automatically producing a summary for the collection C. The only exception was for the summarization of the collection A, in which the produced summary should not be an update summary (as the reader had no previous knowledge). In the more recent TAC conferences, just two text collections were used in each test set (instead of three as in DUC 2007). In general, all these researches and datasets have been carried out for the English language.

For the Portuguese Language, although there are many investigations on traditional Automatic Summarization tasks, (see, e.g., [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16]), there are still few efforts in the US field. In this paper, we have carried out a investigation of main US methods (from distinct approaches) for Portuguese texts, adapting and evaluating them. We also propose a new method by combining the summarization strategies of some other methods, and we achieve better results, advancing the state of the art. More than this, we introduce a reference dataset and establish an experimental environment in order to motivate and foster the investigation of US methods for the Portuguese language.

This paper is organized in the following four sections: we describe the related work in Section II; we present the dataset that was used in this paper in Section III; the experimental setup and the evaluation results are reported in Section IV; and we present some conclusions and final remarks in Section V.

II. Related Work

Researchers have been proposing distinct approaches to produce update summaries, which usually have a source text representation and a method of content selection to compose the summary. In general, most of the works use a sentential representation and a sentence ranking function in order to select the content for the summary. Below, we will present the most representative methods from the different approaches, from the simplest to the more complex ones, and their advantages and disadvantages.

[17] and [18] proposed methods that rank the source sentences based on lexical features to identify clues of updated content. [17] assumes that a good summary must have a word distribution similar to its source texts, and it also shows that the frequencies of words in the old texts may be used to estimate how much outdated the sentences in the new texts are. [18] proposes the Novelty-Factor method, which scores the sentences based on the vocabulary differences among old and new texts using the following equation:

$$NF(s) = \frac{1}{|s|} \sum_{w \in s} \frac{|w \in D_{new}|}{|w \in D_{old}| + |D_{new}|}$$

where s is a sentence, w is a word, and $D_\bullet$ is a collection of $\bullet$ (new or old) documents. As we may see, a sentence s receives a high score when its words occur more times in the texts of a new collection than in an old one.

[19], [20] and [12] use positional features and their results show that this kind of data is better to find salient information than updated information. [19] produces summaries based on the Optimal Position Policy (OPP) rank, which estimates how much relevant a sentence is by its respective position in the text. The authors built the OPP rank by the analysis of the distribution of Elementary Discourse Units (EDUs), as defined in the Pyramid evaluation method [21], for each sentence position in the DUC 2007 dataset. Once the OPP rank is learned, it may be used as a scoring function for the sentences in order to produce the summaries. As it was expected, the selection of first sentences usually produces better results. The authors have referenced to this method as a more robust baseline for update summarization. [12] replicates the experiments with OPP rank for Portuguese language in the CSTNews corpus ([22], [23]). However, the authors use two different information instead of EDUs in order to build the positional rank: the frequencies of words and manually identified sentential alignments among sentences from summaries and their respective source texts in the corpus [24]. It is important to say that the word frequency has been used in a lot of summarization researches and it is very useful to find salient information [25]. [12] shows that the use of sentential alignments produces better results than frequencies of words because they result in a more sophisticated way to identify the content from source texts that was selected to the summary. [20] shows experiments with many positional features of sentences and words, which are based on the idea that the most relevant content occurs first in texts. Thus, their ranking functions decrease the score of a sentence or a word according to their distance to the respective first instance (sentence or word). It is an interesting method because it assumes relevance for first occurrences of words, which may be in other parts of the texts, and it is not limited to the first sentences. They have presented four different positional functions, where n is the number of sentences in the document and i is the position of the sentence that will be ranked, as follows:

- **Direct proportion**: $f(i) = (n - i + 1)/n$;
- **Inverse proportion**: $f(i) = 1/i$;
- **Geometric sequence**: $f(i) = (1/2)^{i-1}$;
- **Binary function**[3]: $f(i) \ = \ 1 \ if \ i == 1 \ else \ \lambda$.

The methods above are simple and fast methods to rank sentences, but they adopt oversimplified text representations that do not identify the information flow among old and new texts in order to find updated content. [17] and [18] analyzed this information, but in a superficial way.

[26] proposes a method based on the differences among LSA [27] topics from old and new texts. Each topic is scored by the subtraction of its weight in old and new texts. Thus, a topic gets a high score if it is more relevant in new texts than others. Iteratively, the best weighted sentence from the topic with the highest score is selected to the summary and the weights are recalculated.

[28] and [29] associate labels (four and three labels, respectively) for LDA topics based on their weights in the old and new texts. As an example, [28] defines the following topics: emergent (topics present only in new texts); active (topics present on both collections, but more relevant in new texts); not active (topics more relevant in old texts); and extinct

[3] [20] has suggested the use of a small positive real number for λ. We have used $\lambda = 0$.

(topics present only in old texts). These methods use different features in order to select the sentences for the summary. [28] uses word frequencies and [29] applies the Maximal Marginal Relevance (MMR) [30] approach, which assumes that a good sentence must be similar to a target and dissimilar to another one, as the new and old texts, respectively. Both first select the sentences related to the topics with higher weights in the new texts.

[1] shows a method based on probabilistic topic models, called DualSum. Each text in this approach is represented by a bag of words and each word is associated with a latent topic similar to the LDA model. DualSum, which has a procedure similar to the TopicSum system [31], learns a distribution of topics that are organized into the following categories: general topic, which works as a language model in order to identify irrelevant information; topics for collections A and B, in which they represent the subjects that are more present in the old and new texts, respectively; and document specific topics. After this learning step, DualSum finds an output update summary with topics closest to a target distribution, which is based on the intuition that a good summary may be more similar to its respective texts in the collection B.

Methods based on graph models have been widely investigated in Automatic Summarization (see, e.g., [32], [33], [34], [35], [10], [36]). To the best of our knowledge, in the context of US, the most expressive results were reached by the Positive and Negative Reinforcement (PNR^2) system [37]. PNR^2 uses a graph for text modeling, in which each node indicates a sentence and each edge between two sentences is weighted by their Cosine similarity [38]. In PNR^2, given a graph that represents a text collection, its procedure runs an optimization algorithm in which the sentences share scores among themselves based on their similarities with positive and negative reinforcements. A positive reinforcement occurs only among sentences from the same text set and it is represented by a positive β parameter in the algorithm. On the other hand, a negative relation occurs among sentences from different sets and it is indicated by a negative α parameter. This way, a sentence receives a high score if it is more similar to sentences from new texts. In the experiments reported in [37], PNR^2 outperforms the PageRank [39] algorithm, which the authors have also experimented for the US task.

III. The corpus for Portuguese: CSTNews-Update

CSTNews-Update is a different setup of the CSTNews corpus [22], [23], which has 50 text collections with two or three related texts that were collected from mainstream news agencies in Brazil. CSTNews has been used in many investigations of Single and Multi-document Summarization methods for the Portuguese language (see, e.g., [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16]). Each text collection in CSTNews is labeled into one of these following categories: daily news, world news, sports, economy, politics, and sciences.

In a similar way to the datasets for US that were used in the TAC conferences, each test set in CSTNews-Update has two collections, A (old) and B (new), that are also chronologically sorted. We want to produce an update summary from the second collection under the assumption that the reader has already read the texts in the first one.

The number of texts in each collection in CSTNews-Update ranges from 1 to 3, and, in each collection A in the corpus, there is only a single text. These restrictions were used in order to build an experimental environment similar to the DUC and TAC conferences, in which there is always more than one text in each collection B.

In CSTNews-Update, there are 58 test sets that were produced based on two different approaches, intra-cluster and inter-cluster. In the first one, all the original text collections in CSTNews that have three texts were used (in a total of 39). In the second approach, pairs of collections that have similar subjects were manually grouped (in a total of 19), forming larger collections.

An interesting feature of CSTNews-Update is its different timestamp distances (from seconds to days) among the texts in the collections A and B for each test set. This feature may model cases of real world, in which the users may read sequential texts that have low timestamp differences and also read others that have huge differences.

As expected, the timestamp differences among documents are low in the intra-cluster collections and huge in the inter-cluster ones. The maximal difference is approximately 216 hours and the average difference is 175.51 hours. Thus, CSTNews-Update enables investigations about the impact of the published time of documents to find updated and new information. However, it is expected that, in the sets with higher timestamp distances among its collections A and B, the identification of the most relevant updated content is harder because probably there are more different information among the texts.

IV. Experimental Setup and Results

In order to evaluate the methods of US that were experimented in this paper, we have applied the ROUGE framework [40], which is the most used evaluation approach in investigations of Automatic Summarization. ROUGE computes the number of n-grams in common among automatic and reference texts (usually, human summaries are used as reference texts), resulting in Precision, Recall and F-measure figures. Their results are indicative of the informativeness of the automatic summaries: the closer to 1 the results are, the more informative the summaries are.

In addition to ROUGE, we have also employed the Nouveau-ROUGE method [41], which is a different application of ROUGE with focus on the US task. This metric assumes that a good update summary must be informative and updated. Thus, initially, Nouveau-ROUGE computes two ROUGE scores, $R^{(AB)}$ and $R^{(BB)}$, in which there are reference texts either from collection A or from collection B respectively. After that, weights are used to approximate the difference between $R^{(AB)}$ and $R^{(BB)}$ to the manual summarization evaluations approaches of Pyramid[4] [21] and

[4] In Pyramid evaluation, automatic summaries are scored by their content units, which are weighted by their occurrences in the reference summaries.

Responsiveness[5], resulting in correlated Precision, Recall and F-measure figures for both of them.

As there are not update summaries made by humans in the CSTNews dataset, we applied the automatic evaluation approach that was proposed in [42], in which the produced summaries are compared to their respective source texts. Here, it is important to say that the experiments that were reported in [42] have shown that the ROUGE evaluation of summaries based on their source texts is a good approximation of evaluations based on human summaries. Therefore, for ROUGE, we compared the produced summaries to their respective source texts that are labeled as collection B. We show the average scores for two of ROUGE running settings, ROUGE-1 and ROUGE-2, which calculate the scores based on unigrams and bigrams overlapping (that are the most used variations), respectively. Furthermore, we applied the same ROUGE parameters[6] that were applied in the TAC conferences.

Once Novelty-ROUGE is focused on the US task and it requires two collections of reference texts (texts in collections A and B), we used the respective old and new texts for each produced summary. Here, we only report the F-measure score and its respective correlation with Responsiveness (Res) and Pyramid (Pyr) results [21].

Following what was observed at the DUC and TAC conferences, we produced update summaries based on the extractive approach, in which the systems pick some sentences from the source texts and put them in the output without content changes. Furthermore, all the produced summaries have no more than 100 words, and it was assumed that the reader had already read the old texts in each document set. However, as we focus in the US task, we present the average evaluation scores for update summaries only, and we do not consider the summaries formed for collection A.

We performed experiments with the most representative methods of the distinct summarization approaches that were investigated in the US task, as follows: DualSum [1], which uses a probabilistic topic model; the graph based algorithms PNR^2 [37] and its variations with distinct setups of the PageRank algorithm [39]; the ranking functions based on positional features that were proposed by [20]; the Novelty-Factor [18] and a method based on the Number of New Words (which is a simplification of the Novelty-Factor, in which only the number of words that occur in the new texts is considered), which are sentential ranking procedures based on vocabulary differences between old and new texts. We have also performed experiments with the RSumm system [10], which is among the best systems for Portuguese for general multi-document summarization, not being tailored for the US task. The purpose of this comparison is to show how (in)adequate such general systems are for the US task.

In our experiments with DualSum, we have used the same setups that were adopted in [1]. Thus, we applied the same preprocessing steps, but changed the resources and tools that are language depended. In the topic learning stage, we use the CSTNews-Update dataset in order to identify the general topics, once [1] has also applied the experimented dataset itself. Here, it is important to say that [1] has proposed that this kind of topic may be previously learned in order to reduce the required computational processing time.

We investigate the PageRank [39] algorithm in two different setups that were also used in [37], the PageRank (A + B) and PageRank (B). In the first one, we use the sentences from the A and B collections in order to build a sentential graph, in which each node is a sentence and each edge indicates the Cosine similarity [38] between two sentences. The second setup has a procedure identical to the first one, but we build the graph with sentences from the collection B only. It is important to notice that, independently of the approach, only sentences from Collection B are used to build the summary. We have also used these two setups in the RSumm method [10], once it is also based on graph algorithms. However, RSumm was not affected by this and produced the same evaluation scores.

Finally, we also propose and test a new method that combines all the previously cited US methods, excepting DualSum. Here, the idea was to investigate the impact of using fast (simple) sentential ranking methods in order to produce update summaries. Thus, each sentence is scored based on the sum of its respective scores in the Novelty-Factor, Number of New Words, PageRank, PNR^2 and Positional features. Here, each individual score from the methods was normalized (from 0.0 to 1.0) and the final weight was a simple sum of them.

For all the above methods, except for DualSum that internally handles content redundancy, we have used the procedure for redundancy removal that was defined in [10]. For each produced summary, we firstly define as threshold t the average Cosine [38] among all the sentences in the collection B[7] and, after that, we ignore those sentences that have similarity above the threshold with any other sentence that has already been included in the output summary. This strategy of varying threshold (that depends on the test set) is interesting because it may better handle different sceneries. For instance, in collections with very similar texts, the used threshold is higher than in contexts in which there are very distinct texts.

Table I shows the evaluation scores of the methods investigated in this paper. We sort the methods by their F-measure scores for ROUGE-2. We organize the methods in the rows and each evaluation score in the columns. For instance, our combined method shows 0.426, 0.329 and 0.561 of F-measure for ROUGE-1, ROUGE-2 and Nouveau-ROUGE related to Pyramid.

Before discussing the results, it is important to be aware of how ROUGE results must be interpreted. As their authors argue, ROUGE is good at comparing automatic summaries, being almost as good as humans in ranking different summaries according to their informativty. Taken in isolation, a single ROUGE value is not a direct indication of the summary quality. More than this, it is important to remember that the summarization task is usually cruel for evaluation, as there is not a unique good summary as reference. Instead, there are many possible summaries for a collection of texts. Such fact

[5] Responsiveness evaluates the "usefulness" of automatic summaries, considering in the evaluation the linguistic quality.

[6] -n 4 -w 1.2 -2 4 -u -c 95 -r 1000 -f A -p 0.5 -t 0 -a

[7] Remember that only sentences from this collection may be added in the summary.

TABLE I
ROUGE AND NOUVEAU-ROUGE SCORES IN THE CSTNEWS-UPDATE DATASET BASED ON THE [42] EVALUATION APPROACH. THE METHODS ARE SORTED BY THE F-MEASURE SCORES OF ROUGE-2.

	ROUGE-1			ROUGE-2			Nouveau-ROUGE	
Methods	Precision	Recall	F-measure	Precision	Recall	F-measure	Res	Pyr
Combined method	0.843	0.296	0.426	0.648	0.230	0.329	3.562	0.561
DualSum	0.823	0.294	0.418	0.636	0.230	0.325	3.343	0.489
Position - Direct	0.831	0.289	0.418	0.632	0.222	0.319	3.413	0.507
Novelty-Factor	0.837	0.288	0.417	0.637	0.221	0.319	3.661	0.562
PNR^{2}	0.814	0.284	0.409	0.628	0.220	0.317	3.466	0.522
PageRank (A+B)	0.817	0.288	0.413	0.620	0.221	0.317	3.241	0.469
PageRank (B)	0.822	0.284	0.411	0.627	0.218	0.314	3.319	0.491
Position - Binary	0.797	0.277	0.400	0.603	0.211	0.304	3.369	0.499
Position - Geometric	0.796	0.277	0.400	0.602	0.211	0.304	3.370	0.499
Position - Inverse	0.795	0.277	0.400	0.601	0.211	0.304	3.362	0.498
New Words	0.765	0.261	0.379	0.561	0.193	0.279	3.675	0.575
RSumm	0.714	0.247	0.356	0.444	0.149	0.217	2.049	0.586

may explain why ROUGE numbers are generally low in the area. Therefore, ROUGE is a comparative measure and must be read and interpreted in this way.

As we may see, the differences among the scores of some methods are not so high. It probably occurs because some test cases in our dataset have short texts or just one text in the new collections, while we have used a fixed summary length, which is the same used in the DUC and TAC conferences. In general, however, we see that some methods significantly outperform others.

Usually, the methods that have more procedures to identify more recent and relevant content present better results, as DualSum, PNR^{2} and Novelty-Factor approaches. It was not expected that the Position-Direct method would outperform many others. However, positional features have been used in many summarization investigations and have presented satisfactory results, being relevant features.

As expected, the RSumm system [10] method does not outperform the others, as it is not focused on US, demonstrating that efforts are needed to tackle the summarization task specificities. It is interesting that the other graph methods have presented better results, as PageRank (A+B) and PageRank (B). This probably happens because they were tailored for the US task.

Overall, we may conclude that our combined method and DualSum were the methods that produced the most informative summaries among the traditional methods we investigated in this paper. Our combined approach was the best one, showing that simple features may be very useful to the task. Regarding the good performance of DualSum, this is not totally surprising, as this approach has constantly achieved good results in the area.

We have also observed that the average ROUGE and Nouveau-ROUGE scores for the investigated methods in the intra-cluster setup is a bit higher than in inter-cluster collections[8]. As we said before, the timestamp differences among the texts that occur in the inter-cluster collections are higher than in the intra-cluster ones, and our results suggest that identifying the most relevant updated content is harder in these situations.

V. Final Remarks

We introduced an experimental setup and a reference dataset, the CSTNews-Update[9] that allow the investigation of Update Summarization methods for the Portuguese language. Based on them, we have evaluated the most representative methods of Update Summarization from different approaches. The evaluation scores show that some performance differences are small, but that some methods significantly outperform others, indicating future research directions for US investigation for Portuguese. In particular, we proposed a new method that aggregates features from other methods, which are simple and fast approaches for sentence selection, achieving the best results in the evaluation.

Particularly, we believe that the US methods might significantly benefit from some more intelligent/linguistically motivated text representation for the collections of previously known/read and new texts, e.g., making use of some kind of semantic representation, including semantic role labels, named entities and concepts.

Future work includes the investigation of deeper and more abstractive (instead of extractive) approaches for the task.

Acknowledgments

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

References

[1] J.-Y. Delort and E. Alfonseca, "DualSum: a topic-model based approach for update summarization," in *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, Avignon, France, 2012, pp. 214–223.

[2] I. Mani, *Automatic Summarization*, R. Mitkov, Ed., 2001, vol. 3.

[3] R. Witte, R. Krestel, and S. Bergler, "Generating update summaries for DUC 2007," in *Proceedings of the Document Understanding Conference*, Rochester, New York USA, 2007, pp. 1–5.

[4] T. A. S. Pardo, "DMSumm: Um gerador automático de sumários," Master's thesis, Universidade Federal de São Carlos, 2002.

[8] We have not included separate tables for such results due to space limitation.

[9] https://github.com/fernandoasevedo/CSTNews-Update

[5] L. H. M. Rino, T. A. S. Pardo, C. N. Silva Jr, C. A. A. Kaestner, and M. Pombo, “A comparison of automatic summarization systems for brazilian portuguese texts,” in *Proceedings of the 17th Brazilian Symposium on Artificial Intelligence*, São Luís, MA, Brazil, 2004, pp. 235–244.

[6] E. Muller, J. Granatyr, and O. R. Lessing, “Comparativo entre o algoritmo de Luhn e o algoritmo GistSumm para sumarização de documentos,” *Revista de Informática Teórica e Aplicada*, vol. 22, no. 1, pp. 584–599, 75–94.

[7] D. S. Leite, L. H. M. Rino, T. A. S. Pardo, and M. das Graças Volpe Nunes, “Extractive automatic summarization: Does more linguistic knowledge make a difference?” in *Proceedings of the HLT/NAACL Workshop on TextGraphs-2: Graph-Based Algorithms for Natural Language Processing*, Rochester, NY, USA, 2007, pp. 17–24.

[8] L. Antiqueira, O. N. Oliveira Jr, L. d. F. Costa, and M. d. G. V. Nunes, “A complex network approach to text summarization,” *Information Sciences*, vol. 179, no. 5, pp. 584–599, 2009.

[9] M. L. Castro Jorge and T. A. S. Pardo, “A generative approach for multi-document summarization using the noisy channel model,” in *Proceedings of the 3rd RST Brazilian Meeting*, Cuiabá, MT, Brazil, 2011, pp. 75–87.

[10] R. Ribaldo, A. T. Akabane, L. H. M. Rino, and T. A. S. Pardo, “Graph–based methods for multi-document summarization: Exploring relationship maps, complex networks and discourse information,” in *Proceedings of the 10th International Conference on Computational Processing of Portuguese (LNAI 7243)*, Coimbra, Portugal, 2012, pp. 260–271.

[11] S. Silveira and A. H. Branco, “Enhancing multi-document summaries with sentence simplification,” in *Proceedings of the 14th International Conference on Artificial Intelligence*, Montreal, Quebec, Canada, 2012, pp. 742–748.

[12] F. A. A. Nóbrega, V. Agostini, R. T. Camargo, A. Di Felippo, and T. A. S. Pardo, “Alignment-based sentence position policy in a news corpus for multi-document summarization,” in *Proceedings of the 11st International Conference on Computational Processing of Portuguese (LNAI 8775)*, São Carlos, SP, Brazil, 2014, pp. 6–9.

[13] P. Cardoso and T. A. S. Pardo, “Joint semantic discourse models for automatic multi-document summarization,” in *Proceedings of the 10th Brazilian Symposium in Information and Human Language Technology*, Natal, RN, Brazil, 2015, pp. 81–90.

[14] M. Ângelo Abrantes Costa and B. Martins, “Uma comparação sistemática de diferentes abordagens para a sumarização automática extrativa de textos em português,” *Linguamática*, vol. 7, no. 1, pp. 23–40, 2015.

[15] P. Cardoso and T. A. S. Pardo, “Multi-document summarization using semantic discourse models,” *Processamiento de Lenguaje Natural*, vol. 56, pp. 57–64, 2016.

[16] R. Ribaldo, P. F. Cardoso, and T. A. S. Pardo, “Exploring the subtopic-based relationship map strategy for multi-document summarization,” *Journal of Theoretical and Applied Computing*, vol. 23, no. 1, pp. 183–211, 2016.

[17] L. H. Reeve and H. Han, “A term frequency distribution approach for the duc-2007 update task,” in *Proceedings of Document Understanding Conference 2007*, Rochester, New York USA, 2007, p. 7.

[18] V. Varma, V. Bharat, S. Kovelamudi, P. Bysani, S. GSK, K. K. N, K. R. K. Kumar, and N. Maganti, “IIIT hyderabad at TAC 2009,” in *Proceedings of the second Text Analysis Conference*, Gaithersburg, Maryland USA, 2009, pp. 1–15.

[19] R. Katragadda, P. Pingali, and V. Varma, “Sentence position revisited: A robust light-weight update summarization baseline algorithm,” in *Proceedings of the Third International Workshop on Cross Lingual Information Access: Addressing the Information Need of Multilingual Societies*, Boulder, Colorado, USA, 2009, pp. 46–52.

[20] Y. Ouyang, W. Li, Q. Lu, and R. Zhang, “A study on position information in document summarization,” in *Proceedings of the 23rd International Conference on Computational Linguistics (Posters)*, Beijing, China, 2010, pp. 919–927.

[21] A. Nenkova and R. Passonneau, “Evaluating content selection in summarization: The pyramid method,” in *Proceedings of HLT-NAACL 2004*, Boston, USA, 2004, pp. 145–152. [Online]. Available: http://acl.ldc.upenn.edu/hlt-naacl2004/main/pdf/91_Paper.pdf

[22] P. Aleixo and T. A. S. Pardo, “*CSTNews: Um Córpus de Textos Jornalísticos Anotados segundo a Teoria Discursiva Multidocumento CST (Cross-document Structure Theory)*,” Instituto de Ciências Matemáticas e de Computação, Tech. Rep. 326, 2008.

[23] P. C. F. Cardoso, E. G. Maziero, M. L. R. Castro Jorge, E. M. R. Seno, A. Di Felippo, L. H. M. Rino, M. d. G. V. Nunes, and T. A. S. Pardo, “CSTNews – a discourse-annotated corpus for single and multi-document summarization of news texts in brazilian portuguese,” in *Anais do III Workshop “A RST e os Estudos do Texto”*, Cuiabá, MT, Brasil, 2011, pp. 88–105.

[24] V. Agostini, R. E. López Condori, and T. A. S. Pardo, “Automatic alignment of news texts and their multi-document summaries: Comparison among methods,” in *Proceedings of the 11st International Conference on Computational Processing of Portuguese*, São Carlos, SP, Brazil, 2014, pp. 220–231.

[25] A. Nenkova and L. Vanderwende, “The impact of frequency on summarization,” Microsoft Research, Tech. Rep., 2005.

[26] J. Steinberger and K. Ježek, “Update summarization based on latent semantic analysis,” in *Text, Speech and Dialogue*, ser. Lecture Notes in Computer Science, V. Matoušek and P. Mautner, Eds., 2009, vol. 5729, pp. 77–84.

[27] T. K. Landauer and S. T. Dutnais, “A solution to plato’s problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge,” *Psychological review*, vol. 104, pp. 211–240, 1997.

[28] L. Huang and Y. He, “Corrrank: Update summarization based on topic correlation analysis,” in *Advanced Intelligent Computing Theories and Applications. With Aspects of Artificial Intelligence*, ser. Lecture Notes in Computer Science, D.-S. Huang, X. Zhang, C. A. Reyes García, and L. Zhang, Eds., 2010, vol. 6216, pp. 641–648.

[29] J. Li, S. Li, X. Wang, Y. Tian, and B. Chang, “Update summarization using a multi-level hierarchical dirichlet process model,” in *Proceedings of the 24th International Conference on Computational Linguistics*, Mumbai, India, 2012, pp. 1603–1618.

[30] J. Carbonell and J. Goldstein, “The use of MMR, diversity-based reranking for reordering documents and producing summaries,” in *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York, NY, USA, 1998, pp. 335–336.

[31] A. Haghighi and L. Vanderwende, “Exploring content models for multi-document summarization,” in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Stroudsburg, PA, USA, 2009, pp. 362–370.

[32] G. Erkan and D. R. Radev, “Lexrank: Graph-based lexical centrality as salience in text summarization,” *Journal of Artificial Intelligence Research*, vol. 22, no. 1, pp. 457–479, 2004.

[33] J. Leskovec, N. Milic-Frayling, and M. Grobelnik, “Extracting summary sentences based on the document semantic graph,” Microsoft Research, Tech. Rep. MSR-TR-2005-07, 2005.

[34] Z. Lin and M.-Y. Kan, “Timestamped graphs: Evolutionary models of text for multi-document summarization,” in *Proceedings of the Second Workshop on TextGraphs: Graph-Based Algorithms for Natural Language Processing*, Rochester, NY, USA, 2007, pp. 25–32.

[35] X. Li, L. Du, and Y.-D. Shen, “Graph-based marginal ranking for update summarization,” in *SDM*, 2011, pp. 486–497.

[36] L. Du, X. Li, , and Y.-D. Shen, “Update summarization via graph-based sentence ranking,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 5, pp. 1162–1174, 2013.

[37] L. Wenjie, W. Furu, L. Qin, and H. Yanxiang, “PNR^2: ranking sentences with positive and negative reinforcement for query-oriented update summarization,” in *Proceedings of the 22nd International Conference on Computational Linguistics*, vol. 1, Stroudsburg, PA, USA, 2008, pp. 489–496.

[38] G. Salton, A. Wong, and C. S. Yang, “A vector space model for automatic indexing,” *Communications of the ACM*, vol. 18, no. 11, pp. 613–620, 1975.

[39] S. Brin and L. Page, “The anatomy of a large-scale hypertextual web search engine,” in *Proceedings of 17th International World-Wide Web Conference*, Brisbane, Australia, 1998, p. 20.

[40] C.-Y. Lin, “ROUGE: A package for automatic evaluation of summaries,” in *Text Summarization Branches Out: Proceedings of the Association Computational Linguistics–04 Workshop*, Barcelona, Spain, 2004, pp. 74–81.

[41] J. M. Conroy, J. D. Schlesinger, and D. P. O’Leary, “Squibs: Nouveau-rouge: A novelty metric for update summarization,” *Computational Linguistics*, vol. 37, no. 1, pp. 1–9, 2011.

[42] A. Louis and A. Nenkova, “Automatically evaluating content selection in summarization without human models,” in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, Singapore, 2009, pp. 306–314.