International Journal of Computational Linguistics and Applications vol. 7, no. 2, 2016, pp. 111–128
Received 08/02/2016, accepted 06/03/2016, final 20/09/2016
ISSN 0976-0962, http://ijcla.bahripublications.com

Improving Content Selection for Update Summarization with Subtopic-Enriched Sentence Ranking Functions

FERNANDO A. A. NÓBREGA AND THIAGO A. S. PARDO

Universidade de São Paulo (USP), Brasil

ABSTRACT

Update Summarization aims to produce summaries under the assumption that the reader had some knowledge about the topic from the source texts. Usually, traditional approaches of summarization use sentential ranking functions in order to find the most relevant and updated sentences from source-texts. We propose the enriching of these methods with the using of subtopic representation, which are coherent textual segments with one or more sentences in a row. The results of our experiments show that our text representation improves the quality of produced summary and show high recall values.

1 INTRODUCTION

Update Summarization aims to produce summaries under the assumption that the reader has some prior knowledge about the topic of the source texts, so that the output summary must show to the reader the most relevant and updated information. This task was introduced at Document Understanding Conference (DUC) of 2007[1], in which for each test case there are 3 text sets (A, B and C) that are

[1] duc.nist.gov/duc2007/tasks.html

sorted by their publication timestamps and a summary for each set should be produced as follows: the first one, is a traditional summary with the most important content from set A; and two update summaries from set B and C so that it is assumed the reader has known the set A and B respectively.

The generation of update summaries requires dealing with a lot of challenges, as: identification of salient information; removal of redundant content; arrangement of the sentences in the summary in order to produce coherent passages; and the analysis of the old (those texts that reader has already read) and new (those texts that reader does not know) sets in order to identify updated information in the new texts, which is the focus of the update summarization.

Previous investigations have proposed the analysis of vocabulary differences over old and new texts in order to identify update content [1, 2]. This approach does not demand much computational power and show satisfactory results. However, once these methods do not analyze textual topics or subjects, they may produce summaries that are not informative as they might be. On the order hand, topic models like Latent Dirichlet Allocation (LDA) [3] or Latent Semantic Analysis (LSA) [4] have been used in order to compute the source texts and to identify the most relevant and update content of them [5–8]. Usually, topic models approaches show very good results, however, they require high computational power and they need to be recomputed when a text is included in a collection (e.g., when a new text is published) [9].

In this paper, aiming to produce more informative update summaries by use methods that require low computational power, we propose the enrichment with textual subject based on subtopic segments into traditional algorithms of content selection for summarization. Here, we follow the definition of [10, 11], in which the main idea of a text can be segment into minor ideas, or its subtopics. So that each subtopic from a text can be represent by a coherent textual segment with one or more sentences in a row. For instance, Table 1 shows a text from the CSTNews corpus [12] with its subtopic segments. In this example, we may see a text about an airplane crash at Democratic Republic of Congo and its main topic (the crash) segmented into three subtopics: sentences from 1 to 5; sentence 6; and sentence 7. The first subtopic is about the accident itself and the others present more details about the airplane and its crew, respectively.

Table 1. Example of a text segmented into subtopics

$[S_1]$	A plane crash in Bukavu, in the Eastern Democratic Republic of Congo, killed 17 people on Thursday afternoon, said the spokesman of the United Nations.
$[S_2]$	The victims of the accident were 14 passengers and three crew members.
$[S_3]$	Everyone died when the plane, hampered by the bad weather, failed to reach the runway and crashed in a forest that was 15 kilometers from the airport in Bukavu.
$[S_4]$	The plane exploded and caught fire, said the UN spokesman in Kinshasa, Jean-Tobias Okala.
$[S_5]$	“There were no survivors”, said Okala.
$[S_6]$	The spokesman said the plane, a Soviet Antonov-28, of Ukrainian manufacturing and under ownership of the Trasept Congo, a Congolese company, also took a mineral load.
$[S7]$	According to airport sources, the crew members were Russian.

It is important to say once it is possible to identify the subtopic segments of a text by use a process with low computation cost, as the TextTiling algorithm [10], we can add this subtopic representation into many summarization approaches, which usually apply some kind of sentential ranking in order to pick those sentences that must be included in the output summary, without taking much more time for the preprocessing of source texts. Thus, we have incorporated subtopic information into three content selection methods for summarization based on sentential ranking functions that use different metrics, as: vocabulary differences over old and new texts; positional features; and Maximal Marginal Relevance (MMR) [13].

The main motivation for this study is that content selection may be better if we consider that update summaries should mainly contain subtopics from the new texts that the reader has not seen before. Our experiments show that the subtopic-enriched versions of these functions produce summaries with some informativeness gain.

The rest of this paper is organized as follows. We introduce the main related work in Section 2. In Section 3, we present our approach to enrich the sentence ranking functions and in Section 4 we propose a sentential ranking with subtopics based on two steps. In Section 5, we describe the data set we used in this work and the setup of our experiments. We show the evaluation results in Section 6. Some final remarks are presented in Section 7.

2 Related Work

Researchers have proposed distinct approaches to produce update summaries, which usually have a textual representation and a method of content selection. In general, most of them use a sentential representation and a ranking function in order to select the content for the summary. Below, we will introduce some relevant methods, from the simplest to the more complex ones, and their advantages and disadvantages.

[1] and [2] use sentence ranking functions based on lexical features in order to find updated information. [1] assumes that a good summary must have a word distribution similar to its source texts, showing that the frequencies of words in the old texts may be used to estimate how much outdated the sentences in the new texts are. [2] proposes the Novelty Factor method that analyzes the vocabulary differences among old and new texts by a simple mathematical formulation. These methods have simple ranking functions and show good results, however, they do not model the text subtopics; they only use bag of words and look at individual features of words in the texts.

[14] and [15] use positional features and their results indicate that this kind of data is better to find salient information instead of updated information. [14] produces summaries based on the Optimal Position Policy (OPP) ranking that estimates how relevant a

sentence is by its position in the text. The authors produce the OPP ranking by the analysis of the distribution of Elementary Discourse Units (EDUs), as defined in the Pyramid evaluation method [16], for each sentence position in the DUC 2007 data set. The authors have referenced this method as a more robust baseline for update summarization. [15] shows experiments with many positional features for sentences and words, which are based on the idea that the most relevant content occurs first in texts. The 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 considers relevance for first occurrences of words, which may be in other parts of the texts, and it is not limited to the first sentences.

The methods above present simple and fast methods to rank sentences, however, they use relatively simple representation of texts that do not identify the information flow among old and new texts in order to find updated content. In order to represent the relations between old and new information in a better way, some authors present methods based on topic models, as LDA [3] or LSA [4].

[5] proposes a method based on the differences among LSA topics from old and new texts, in which 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.

[6] and [8] associate labels for LDA topics based on their weights in the old and new texts. For instance, [6] 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 (topics present only in old texts). These methods use different features in order to select the sentences for the summary. [6] uses word frequencies and [8] applies the Maximal Marginal Relevance (MMR) [13] ap-

proach. Both approaches select first the sentences related to topics with high weight in new texts.

[7] 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. The topics are scored by their relevance for each single text, new and old texts individually, and all the texts. DualSum uses a probabilistic model to find an update summary with topic distribution closest to the new texts.

In general, these topic model approaches estimate the text subjects and their distribution in the source texts. This way, they may capture many rich relations among the sentences and produce more informative summaries. However, as presented by [9], they require high computational power and they need to be recomputed when a text is included in a collection. Thus, these approaches may not be indicated for very dynamic situations, when new texts are frequently produced (as it happens in the web). In what follows, we detail our approach to consider subtopic information for update summarization.

3 Our Approach

[17] suggests the generation of summaries automatically requires three main steps, as follows: textual analysis, in which the source texts are represented in same computational model in order to be processed; content transformation, in which the summarization methods identify the content that must be included into the output summary; and, synthesis, in which the output summary is finally produced. Furthermore, this process performs until a given compression rate is reached, as a given number of words.

In the textual analysis step, we normalize[2] all the texts (tokenization, removal of stopwords, and stemming) and identify their subtopic segments. In the transformation step, we perform some

[2] We have used available tools in the NLTK package, available at: `http://www.nltk.org/`

sentential ranking function that was enriched with subtopic segments in order to identify the most relevant sentences from the source texts. Finally, we use the Extractive approach (it is the most used synthesis process for summarization), which just organizes the picked sentences into the output summary (we do not produce new sentences or change those that were picked).

We use the TextTiling algorithm [10] for identification of subtopics if the source texts. This method analysis vocabulary differences over each pair of sentences that occur side by side in a given text in order to identify subject boundaries. So that there is a boundary between two sentences s_1 and s_2 if and just if they occur in different subtopics.

Other methods for subtopic identification have been proposed, as [18], which have investigated linguist knowledge based on discursive theories in order to improve the quality of the TextTilling algorithm, and [19], which have proposed a subtopic segmentation based on the LDA model. However, once the traditional TextTiling shows satisfactory results and it is relatively faster than these other approaches, we have picked this one in order to not increase overmuch the running time of the summarization methods that were investigated in this paper, which usually produce summaries very quickly.

In order to reduce the summary redundancy, we avoid candidate sentences (that were picked in the transformation step) that are very similar to some sentence that has been already included into the summary. We use the Cosine metric [20] as sentential similarity score and we define a sentence is similar to another one whether the similarity value between them is higher than a given threshold. As similarity threshold, we use the [21] approach for Multi-document Summarization, in which the threshold is dynamically defined based on the maximal and minimal similarity scores among all the sentences of a text (or a text collection) divided by 2 ($\frac{max-min}{2}$). This way, we have a dynamic value to identify similar sentences on different summarization situations. For instance, if there are many similar sentences in the source texts, the threshold will be higher than in situations with very distinct texts.

We incorporated subtopic information into three ranking functions based on the following metrics: Maximal Marginal Relevance (MMR) [13]; Novelty Factor [2]; and the positional features that were proposed in [15]. We have chosen these methods because they show satisfactory results and low computational cost, and they are based on different summarization approaches (sentential similarity, vocabulary and positional features). In the next subsections, we will explain how we add subtopic information in each of these methods. Furthermore, we also propose a content selection based on two-step ranking process, which we will show in the Section 4.

3.1 *

Maximal Marginal Relevance

The Maximal Marginal Relevance (MMR) formulation has been used for Query Focused tasks as Information Retrieval and also for Automatic Summarization as well. For instance, [22] proposed an adaptation of MMR for ranking sentences of a collection based on their similarities to a given user query and also to sentences from old texts in order to produce update summaries.

Once we have experimented and evaluated different summarization methods that are not query based, we use the MMR approach in order to rank the sentences by their similarities to sentences of the sets with old and new texts, as we may see in the Equation 1, where: s is the sentence will get a score; D_C represents all the sentences in the collection C (old or new), excluding the sentence s; and α is an algorithm parameter to weight how important the new sentences are in relation to the old ones. Thus, a sentence gets a high score when it is more similar to sentences of new texts than others of old texts.

$$\begin{aligned} mmr(s) = \alpha \max_{s_j \in D_{new}} (cosine(s, s_j)) \\ - (1 - \alpha) \max_{s_j \in D_{old}} (cosine(s, s_j)) \end{aligned} \tag{1}$$

Based on the equation above, we propose an adapted MMR for ranking subtopic segments, as we may see in the Equation 2, where: sub is a subtopic; and $s \in sub$ represents all the sentences from the subtopic sub. Here, we rank a subtopic based on the similarity of its sentences to others of old and new texts.

$$\begin{aligned} mmr(sub) = \alpha \max_{s \in sub}[\max_{s_j \in D_{new}}(cosine(s, s_j))] \\ - (1-\alpha)\max_{s \in sub}[\max_{s_j \in D_{old}}(cosine(s, s_j))] \end{aligned} \quad (2)$$

Finally, given the two MMR formulations above, we score each sentence s as follows: $score(s) = mmr(s) + mmr(sub_s)$, where sub_s is the subtopic where s occurs. This way, we may identify the most relevant sentence based on its MMR score and also with the relevancy (based on MMR) of its respective subtopic. It is important to say that we have experimented different values for α and the best results in our experiments were obtained with $\alpha = 0.7$.

3.2 *

Novelty Factor

Novelty Factor (NF) is a sentential ranking function based on lexical scores that are calculated over the vocabulary differences among old and new texts. NF ranks each sentence by the scores of its words and it also normalizes this rank by the respective sentence size (number of words) in order to avoid a formulation bias for sentences with more words. We may see the NF formulation in the Equation 3, where: s is a sentence, w is a word; $|w \in D_C|$ indicates the number of documents of set C (new or old) where the word w occurs. As we may see, the score for a word w is reduced if it has high frequency in the old texts.

$$NF(s) = \frac{1}{|s|}\sum_{w \in s}\frac{|w \in D_{new}|}{|w \in D_{old}| + |D_{new}|} \quad (3)$$

We propose a Novelty Factor based on subtopic segments instead of documents, as we may see in the Equation 4, where: Sub_C represents the subtopics that occur in the set of texts C (new or old). This way, a sentence has a high score if its words occur more in subtopics of new texts than of old texts. Furthermore, even two different words occur in the same texts, they can have different scores because they can occur in different subtopics. In other words, our NF formulation considers the vocabulary differences over textual subjects (that are represented as subtopic segments) instead of documents.

$$NF_{sub}(s) = \frac{1}{|s|} \sum_{w \in s} \frac{|w \in Sub_{new}|}{|w \in Sub_{old}| + |Sub_{new}|} \qquad (4)$$

3.3 *

Positional Features

[15] proposed four different sentential ranking functions based on positional features of words and/or sentences. Basically, each function assumes higher scores for the first occurrence of a word or sentence and decreases these scores for the next occurrences in different scales, as below, where: i is the position of an element (word or sentence) in a document; and n is the number of elements in a document:

- **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]: $f(i) \;=\; 1\, if\; i == 1\; else\; \lambda$.

In order to add subtopic information for the positional features above, we rank a sentence by the sum of its positional score with the score of its respective subtopic. Here, we also score the subtopics by some positional features above. However, in this case, the

[3] [15] has suggested the use of a small positive real number for λ. We have used $\lambda = 0$.

argument i indicates the subtopic position in the document and n represents the number of subtopics in the text.

4 A Two-Step Ranking Process Based on Subtopics

As we may see in the previous sections, we use subtopic information with traditional methods in order to select the sentences for the summary in a single sentential ranking step. This way, we just use the subtopic segments of a text as a background context for sentential ranking. However, based on the idea we may use the subtopic information in order to approximate or estimate the text content, once $[10, 11]$ has defined subtopic as a portion of the main idea of a text, we propose a two-step ranking process, in which we firstly identify the most salient subtopics and, after that, we pick their most relevant sentences for the summary.

In order to rank the subtopics, we use Equation 5, where: $|sub|$ is the number of sentences in the subtopic sub; and f is a sentential ranking function with subtopic information that we have proposed in this paper.

$$subtopic\ score(sub) = \frac{1}{|sub|} \sum_{s \in sub} f(s) \tag{5}$$

In each iteration, we pick a candidate sentence from the best ranked subtopic and include it in the summary (if it is not redundant). Then, we remove the picked sentence form is respective subtopic and we recalculate the subtopic ranking. This way, we reduce the weight of the last selected subtopic and its respective probability to be the next selected one, improving the summary recall by choosing other subtopics.

5 Dataset and Experimental Setup

We use the data set for Update Summarization of the DUC 2007 conference. In this corpus, there are 10 different collections with news texts in English language. In each collection, there are 3 sets of related texts, A, B and C that are sorted by timestamps (time(A) $<$ time(B) $<$ time(C)).

We use the ROUGE [23] system, which is the most used evaluate approach for Update Summarization, in order to evaluate the informativeness of the produced summaries. ROUGE compares the produced summaries with reference texts, which usually are summaries made by humans, based on the analysis of n-grams overlapping. We will show the values of Precision, Recall and F-measure for two settings of ROUGE, ROUGE-1 (for one-grams overlapping) and ROUGE-2 (for bi-grams), with the same parameters that were used in the DUC conference[4].

We only produced and evaluated summaries with no more than 100 words, which is the same limit of summary length that was used at DUC 2007, for each text collection in the data set for the following situations: (i) the reader requires an update summary of set B given he has already read the set A, and (ii) the reader requires an update summary of set C given that set B was read before. We did not produce summaries for set A of the text collections because the focus of this study was the update summarization process only.

6 Evaluation and Results

Tables 2 and 3 show the ROUGE values for the summarization methods based on sentential ranking approaches that were experimented in this paper. We use “sub" and “TwoSteps" in order to label the methods that incorporate subtopic segments and/or our ranking approach based on two steps, respectively. For instance, the caption “MMR + TwoSteps + sub” indicates the results of the MMR approach that was enriched with subtopic segments and that also uses our two steps ranking.

We grouped the results for each summarization approach in order to show a better visualization of the impact of enrichment with some subtopics for each one of them. Inside each group, we sorted the methods by their respective F-measure of ROUGE-2 results and

[4] Parameters for ROUGE: -n 4 -w 1.2 -m -2 4 -u -c 95 -r 1000 -f A -p 0.5 -t 0 -a.

we highlight the highest scores in bold for the other metrics. Furthermore, we also show the three best classified systems at DUC 2007 conference.

Table 2. Summary informativeness evaluation results, part 1

	ROUGE-1		
Methods	**P**	**R**	**F**
#1 duc id 40	**0.374**	0.370	**0.371**
#2 duc id 45	0.344	0.357	0.350
#3 duc id 44	0.361	0.376	0.368
MMR+ sub	**0.340**	0.383	**0.360**
MMR + TwoSteps + sub	0.330	0.376	0.350
MMR	0.309	**0.401**	0.346
NF + TwoSteps + sub	0.320	0.371	0.343
NF + sub	0.321	**0.379**	**0.346**
NF	**0.321**	0.371	0.343
Pos_Geometric+ sub	0.310	**0.378**	**0.337**
Pos_Inverse+ sub	0.310	**0.378**	**0.337**
Pos_Binary	0.310	0.365	0.332
Pos_Binary+ sub	**0.311**	0.367	0.332
Pos_Direct	0.308	0.367	0.331
Pos_Geometric	0.308	0.367	0.331
Pos_Inverse	0.308	0.367	0.331
Pos_Direct+ sub	0.308	0.368	0.332

As one may see, most of the subtopic versions of the methods systematically show slightly better results than their respective original versions. For instance, MMR + sub shows 0.3605 and 0.0798 F-measure values for ROUGE-1 and ROUGE-2, respectively (against 0.3469 and 0.0752 values for the original MMR approaches). Overall, looking at ROUGE-2, the best method was MMR with subtopic information, but the differences to the other subtopic-enriched versions are minimal.

Our best method – MMR with subtopic information – achieved a 0.0798 F-measure for ROUGE-2, which is not very far from the third system in DUC 2007 conference. Furthermore, looking at re-

Table 3. Summary informativeness evaluation results, part 2

	ROUGE-2		
Methods	**P**	**R**	**F**
#1 duc id 40	**0.111**	**0.111**	**0.111**
#2 duc id 45	0.092	0.096	0.093
#3 duc id 44	0.089	0.093	0.091
MMR+ sub	**0.075**	**0.085**	**0.079**
MMR + TwoSteps + sub	0.071	0.081	0.076
MMR	0.067	0.085	0.075
NF + TwoSteps + sub	0.074	0.086	0.080
NF + sub	**0.072**	**0.086**	**0.078**
NF	0.072	0.083	0.077
Pos_Geometric+ sub	0.074	**0.086**	**0.079**
Pos_Inverse+ sub	0.074	**0.086**	**0.079**
Pos_Binary	**0.074**	0.084	0.079
Pos_Binary+ sub	**0.074**	0.084	0.079
Pos_Direct	0.073	0.083	0.078
Pos_Geometric	0.073	0.083	0.078
Pos_Inverse	0.073	0.083	0.078
Pos_Direct+ sub	0.073	0.084	0.078

call for ROUGE-1, it is possible to see that our method outperformed the best systems.

As you may see, the NF + TwoSteps + sub and also the MMR + TwoSteps + sub methods show good results. Regarding the values of Recall for ROUGE-1, they respectively show the third (0.3717) and second (0.3764) best values. However, only the Novelty Factor with our two-step ranking approach showed better results than their other versions, Novelty Factor and Novelty Factor + sub. It is important to say we do not show the results of the methods based on Positional Features and subtopics with our two step ranking approach because of the were not differences among the summaries produces by them to those produced by the one step sentential ranking.

7 FINAL REMARKS

In this paper, we have investigated the use of subtopic information to enrich update summarization approaches, in order to achieve better results at a low computational cost. In our experiments, we may see that our approach slightly improves the informativeness of the summaries produced by some traditional sentence ranking functions. We have also presented a two-step ranking approach, in which we rank the subtopics and then their respective sentences, which have shown some tiny improvements over Recall scores of ROUGE.

Although the performance differences are not very high, it is interesting to notice that they are simple to achieve and may be useful for dynamic situations, in which there are many texts and new texts are quickly produced and made available, as usually happens in the web.

Interestingly, in other experiments that we performed, clustering the subtopic segments did not improve the results and, for this reason, we have not reported these experiments in this paper. Subtopic clustering may be carried out because it is common that the same subtopics in a source text are repeated in the other texts. However, the sentence ranking functions that we tested were not affected by this.

As future work, we envision to try other strategies based on topic/subtopic information, for instance, to use subtopic information with topic model approaches as DualSum [7].

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

REFERENCES

1. Reeve, L.H., Han, H.: A term frequency distribution approach for the duc-2007 update task. In: Proceedings of DUC 2007 (online), Rochester, New York USA (2007) 7

2. Varma, V., Bharat, V., Kovelamudi, S., Bysani, P., GSK, S., N, K.K., Kumar, K.R.K., Maganti, N.: IIIT hyderabad at TAC 2009. In: Proceedings of the second TAC, Gaithersburg, Maryland USA (2009) 1–15
3. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3** (2003) 993–1022
4. Landauer, T.K., Dutnais, S.T.: A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological review **104** (1997) 211–240
5. Steinberger, J., Ježek, K.: Update summarization based on latent semantic analysis. In Matoušek, V., Mautner, P., eds.: Text, Speech and Dialogue. Volume 5729 of Lecture Notes in Computer Science. Springer Berlin, Heidelberg (2009) 77–84
6. Huang, L., He, Y.: Corrrank: Update summarization based on topic correlation analysis. In Huang, D.S., Zhang, X., Reyes García, CarlosAlbertoand Zhang, L., eds.: Advanced Intelligent Computing Theories and Applications. With Aspectsof Artificial Intelligence. Volume 6216 of Lecture Notes in Computer Science. Springer Berlin Heidelberg (2010) 641–648
7. Delort, J.Y., Alfonseca, E.: DualSum: a topic-model based approach for update summarization. In: Proceedings of the 13th Conference of the EACL, Avignon, France, Association for Computational Linguistics (2012) 214–223
8. Li, J., Li, S., Wang, X., Tian, Y., Chang, B.: Update summarization using a multi-level hierarchical dirichlet process model. In: Proceedings of the 24th COLING, Mumbai, India (2012) 1603–1618
9. Wang, D., Li, T.: Document update summarization using incremental hierarchical clustering. In: Proceedings of the 19th ACM CIKM, New York, NY, USA, ACM (2010) 279–288
10. Hearst, M.A.: Texttiling: Segmenting text into multi-paragraph subtopic passages. Comput. Linguist. **23**(1) (1997) 33–64
11. Koch, I.: Introdução à linguística textual. Contexto (2009)
12. Cardoso, P.C.F., Maziero, E.G., Castro Jorge, M.L.R., Seno, E.M.R., Di Felippo, A., Rino, L.H.M., Nunes, M.d.G.V., Pardo, T.A.S.: 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, Sociedade Brasileira de Computação (2011) 88–105
13. Carbonell, J., Goldstein, J.: The use of MMR, diversity-based reranking for reordering documentsand producing summaries. In: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '98, New York, NY, USA, Association for Computing Machinery (1998) 335–336
14. Katragadda, R., Pingali, P., Varma, V.: 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 (CLIAWS3), Boulder, Colorado, Association for Computational Linguistics (2009) 46–52

15. Ouyang, Y., Li, W., Lu, Q., Zhang, R.: A study on position information in document summarization. In: Proceedings of the 23rd CICLING (Posters). COLING '10, Stroudsburg, PA, USA, Association for Computational Linguistics (2010) 919–927
16. Nenkova, A., Passonneau, R.: Evaluating content selection in summarization: The pyramid method. In: Proceedings of HLT-NAACL 2004, Boston, USA, Association for Computational Linguistics (2004) 145–152
17. Mani, I.: Automatic Summarization. Volume 3. John Benjamins Publishing Company (2001)
18. Cardoso, P.C.F., Taboada, M., Pardo, T.A.S.: On the contribution of discourse structure to topic segmentation. In: Proceedings of the 14th SIGDIAL, Metz, France (2013) 92–96
19. Riedl, M., Biemann, C.: TopicTiling: a text segmentation algorithm based on lda. In: Proceedings of ACL 2012 Student Research Workshop, Jeju Island, Korea, Association for Computational Linguistics (2012) 37–42
20. Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic indexing. Communications of the ACM **18**(11) (1975) 613–620
21. Ribaldo, R., Akabane, A.T., Rino, L.H.M., Pardo, T.A.S.: Graph–based methods for multi-document summarization: Exploring relationship maps, complex networks and discourse information. In: Proceedings of the 10th PROPOR (LNAI 7243), Coimbra, Portugal (2012) 260–271
22. Boudin, F., El-Bèze, M., Moreno, J.M.T.: A scalable MMR approach to sentence scoring for multi-documentupdate summarization. In: Proceedings of the 20th COLING (Posters and Demonstrations), Manchester, UK (2008) 23–26
23. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text Summarization Branches Out: Proceedings of the ACL-04 Workshop, Barcelona, Spain (2004) 74–81

FERNANDO A. A. NÓBREGA
NÚCLEO INTERINSTITUCIONAL DE LINGUÍSTICA
COMPUTACIONAL (NILC),
INSTITUTO DE CIÊNCIAS MATEMÁTICAS E DE COMPUTAÇÃO
(ICMC),
UNIVERSIDADE DE SÃO PAULO (USP),
CAIXA POSTAL 668, 13560-970, SÃO CARLOS (SP), BRASIL
E-MAIL: <FASEVEDO@ICMC.USP.BR>

THIAGO A. S. PARDO
NÚCLEO INTERINSTITUCIONAL DE LINGUÍSTICA COMPUTACIONAL (NILC),
INSTITUTO DE CIÊNCIAS MATEMÁTICAS E DE COMPUTAÇÃO (ICMC),
UNIVERSIDADE DE SÃO PAULO (USP),
CAIXA POSTAL 668, 13560-970, SÃO CARLOS (SP), BRASIL
E-MAIL: <TASPARDO@ICMC.USP.BR>