Subtopic Annotation in a Corpus of News Texts: Steps Towards Automatic Subtopic Segmentation

Paula C. F. Cardoso[1], Maite Taboada[2], Thiago A. S. Pardo[1]

[1]Núcleo Interinstitucional de Linguística Computacional (NILC)
Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo
Av. Trabalhador São-carlense, 400 – Centro
Caixa Postal: 668 – CEP: 13566-970 – São Carlos/SP

[2]Department of Linguistics – Simon Fraser University
8888 University Dr., Burnaby, B.C., V5A 1S6 - Canada

pcardoso@icmc.usp.br, mtaboada@sfu.ca, taspardo@icmc.usp.br

Abstract. *Subtopic segmentation aims at finding the boundaries among text passages that represent different subtopics, which usually develop a main topic in a text. Being capable of automatically detecting subtopics is very useful for several Natural Language Processing applications. This paper describes subtopic annotation in a corpus of news texts written in Brazilian Portuguese. In particular, we focus on answering the main scientific questions regarding corpus annotation, aiming at both discussing and dealing with important annotation decisions and making available a reference corpus for research on subtopic structuring and segmentation.*

Resumo. *Segmentação topical visa segmentar um texto em passagens que representam subtópicos diferentes, os quais desenvolvem um tópico principal de um texto. A identificação de subtópicos é útil para diversas aplicações de Processamento de Linguagem Natural. Este artigo descreve a anotação de subtópicos em um córpus de textos jornalísticos em Português do Brasil. Em particular, foca-se em responder as questões cientificas a respeito da anotação do córpus, visando discutir e lidar com questões importantes de anotação e disponibilização de um córpus de referência para pesquisas sobre estruturação e segmentação topical.*

1. Introduction

Subtopic segmentation aims at finding the boundaries among text passages that represent different subtopics, which usually develop the main topic in a text. For example, a text about a football match would have "football" as the main topic and passages that might represent the subtopics "preparation and training for the match", "moves of the match and final score", and "schedule of next matches".

This task is useful for many important applications in Natural Language Processing (NLP), such as automatic summarization, question answering, and information retrieval and extraction. For instance, Prince and Labadie (2007) explain

Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology, pages 49–58,
Fortaleza, CE, Brazil, October 21–23, 2013.

that information retrieval with the identification of subtopics in the retrieved texts may provide the user with text fragments that are semantically and topically related to a given query. This makes it easier for the user to quickly find the information of interest. Oh et al. (2007) suggest that a question answering system, which aims to answer a question/query submitted by the user, may link this query to the subtopics in a text in order to increase the accuracy of the identification of the answer. Wan (2008) says that, given some subtopic segmentation, automatic summarization may produce summaries that select different aspects from the collection of texts, producing better summaries.

Given its usefulness, it is common to prepare a reference segmentation that supports not only the study and understanding of the phenomenon, but also the development and evaluation of systems for automatic subtopic segmentation. As the construction of corpora is a time consuming and very expensive task, it is necessary to follow procedures to systematize it and to ensure a reliable annotation, in order to produce a scientifically sound resource. Hovy and Lavid (2010) claim that it is necessary to be concerned with the reliability, validity, and consistency of the corpus annotation process. Because of this, many researchers, with such procedure in mind, suggest some methodological research questions, which may be summarized in the following 7 steps: (1) choosing the phenomenon to annotate and the underlying theory, (2) selecting the appropriate corpus, (3) selecting and training the annotators, (4) specifying the annotation procedure, (5) designing the annotation interface, (6) choosing and applying the evaluation measures, and (7) delivering and maintaining the product.

In this paper, we report the subtopic annotation of a corpus of news texts written in Brazilian Portuguese. The corpus, called CSTNews[1] (Cardoso et al., 2011), was originally designed for multi-document processing and contains 50 clusters, with each cluster having 2 or 3 texts on the same topic. In particular, we focus on answering the 7 research questions cited above, aiming at both discussing and dealing with important annotation decisions and making available a reference corpus for research on automatic subtopic structuring and segmentation.

The remainder of this paper is organized into two main sections. Section 2 shows a brief discussion about relevant issues related to corpus annotation and some work on subtopic segmentation. Section 3 describes our annotation under the light of the 7 annotation questions, including the description of our dataset and the quality evaluation of the subtopic segmentation. Section 4 presents final remarks.

2. Related work

There are several initiatives to create corpora that are linguistically annotated with varied phenomena from diverse perspectives, both for written and for spoken/transcribed data. We briefly overview some of these works in what follows.

Hearst (1997) was one of the pioneer works in the area of subtopic segmentation with the proposal of the TextTiling algorithm. The author used a corpus of 12 magazine (expository) articles that had their subtopics segmented by technical researchers. The size of the texts varied from 1,800 to 2,500 words. In order to produce a reference

[1] CSTNews corpus - http://www2.icmc.usp.br/~taspardo/sucinto/cstnews.html

segmentation, Hearst considered that a boundary was true if at least three out of the seven enrolled judges placed a boundary mark there. Kazantseva and Szpakowicz (2012), in turn, chose a fiction book (with 20 chapters) to be segmented by at least six undergraduate students. For each topic boundary, the annotator provided a brief one-sentence description, effectively creating a chapter outline. In these two studies, texts were segmented based on paragraph boundaries. The agreement among annotators was 0.64 in Hearst's annotation and 0.29 in Kazantseva and Szpakowicz's study.

Passonneau and Litman (1997) used a corpus composed of 20 transcribed narratives about a movie to be segmented by seven untrained annotators. The size of the narratives was roughly 13,500 words. The authors requested judges to mark boundaries using their notion of communicative intention as the segmentation criterion. Galley et al. (2003) worked on a sample of 25 meetings transcribed from the ICSI Meeting corpus (Janin et al., 2003). They had at least three human judges to mark each speaker change (which is a potential boundary) as either a boundary or non-boundary. The final segmentation was based on the opinion of the majority. Gruenstein et al. (2007) used 40 meetings from the same corpus and 16 additional ones from the ISL Meeting Corpus (Burger et al., 2002). The authors asked to two annotators to segment the texts at two levels: major and minor, corresponding to the more and less important topic shifts. The authors noticed many cases where topic boundaries were annotated as a major shift by one annotator and as a minor shift by other, suggesting low agreement. Mohri et al. (2010) used 447 news of the TDT corpus of broadcast news speech and newspaper articles. The corpus has human-labeled story boundaries treated as topic boundaries.

Other researchers automatically produce reference segmentation. Choi (2000) produced an artificial test corpus of 700 documents from the Brown corpus. For document generation, the procedure consists extracting, for instance, ten segments of 3-11 sentences each, taken from different documents and combining them to form one document. Chang and Lee (2003), in turn, collected 1285 writings to be segmented into subtopics using their method for topic segmentation.

It is interesting to notice how varied the annotation procedures in the above works were. This is expected, since corpora are created for several different (linguistic and computational) purposes. However, corpus annotation practices have evolved with time and some basic steps are expected to be followed in the research. As already cited in the first section, Hovy and Lavid (2010) split the corpus annotation in seven steps. The authors claim that it is necessary to follow these steps in order to have reliable corpora and, therefore, trustworthy applications. In what follows, we present our corpus annotation following the 7 steps proposed by Hovy and Lavid.

3. Subtopic annotation in the CSTNews corpus

3.1. The decision on the linguistic phenomenon to annotate

The phenomenon to be investigated is subtopic segmentation in news texts. Koch (2009) describes that a text can be considered coherent if it displays continuity, i.e., the topical progression must take place so that there are no breaks or interruptions overly long on topic in progress. In this work, following Hearst (1997), we assume that a text or a set of texts develop a main topic, exposing several subtopics as well. We also assume that a topic is a particular subject that we write about or discuss (Hovy, 2009), and subtopics

are represented in pieces of text that cover different aspects of the main topic (Hearst, 1997; Hennig, 2009). News texts usually do not have explicit marking of subtopics; however, the topicality exists as an organizing principle of the text. As an example, Figure 1 shows a short text (translated from one of the texts in CSTNews corpus) with sentences identified by numbers between square brackets and a possible segmentation. We also show the identification of each subtopic in angle brackets after the corresponding text passages. The main topic is a plane crash. The first text block is about the victims and where the plane was; the second block describes the plane; and the last one describes the crew.

[S1] 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.
[S2] The victims of the accident were 14 passengers and three crew members.
[S3] All 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.
[S4] The plane exploded and caught fire, said the UN spokesman in Kinshasa, Jean-Tobias Okala.
[S5] "There were no survivors", said Okala.
<subtopic: plane crash in the Congo>

[S6] The spokesman said the plane, a Soviet Antonov-28 and Ukrainian manufacturing and ownership of the Trasept Congo, a Congolese company, also took a mineral load.
<subtopic: details about the plane >

[S7] According to airport sources, the crew members were Russian.
<subtopic: details about the flight crew>

Figure 1. Example of a text with identified subtopics

It is usual to find subtopics that are repeated later on in the same text, and should therefore be connected, and marked with the same label. Another important matter is that the granularity of a subtopic is not defined, as a subtopic may contain one or more sentences or paragraphs. Some researchers use paragraphs as the basic information unit (e.g., Hearst, 1997; Kazantseva and Szpakowicz, 2012), while others employ sentences (e.g., Chang and Lee, 2003; Riedl and Biemann, 2012).

We also do not distinguish among the notions of subtopic shift and subtopic drift, as proposed by Carlson and Marcu (2001). According to Carlson and Marcu, a subtopic shift is a sharp change in focus, while a drift is a smooth change from the information presented in the first span to the information presented in the second. In our annotation, they were both classified as a change in subtopic.

3.2. Selecting the corpus

We used a corpus composed of 50 clusters of news texts written in Brazilian Portuguese, collected from several sections of mainstream news agencies in Brazil: Politics, Sports, World, Daily News, Money, and Science. The corpus contains 140 texts altogether, amounting to 2,088 sentences and 47,240 words. On average, the corpus contains, for each cluster, two or three texts, and 41.76 sentences and 944.8 words, on average.

The choice for news texts was due to our interest in pursuing general purpose subtopic segmentation. News texts make use of everyday language and are widely

accessible by ordinary people. Such texts usually do not present structural demarcation (as the sections of a scientific paper, for instance), causing automatic subtopic segmentation to be more desirable for NLP tasks and equally challenging.

Each cluster contains texts on the same topic as the corpus was built to foster research in multi-document processing, mainly in multi-document summarization. It is interesting to notice that, since the selection of texts to compose the corpus was driven by which topics were current at the time (because current topics are likely to be commented and covered by different news sources), the distribution of texts is not uniform among sections and agencies. For instance, some sections, as World and Daily News, have far more texts than Science and Money sections. We consider that such differences are not relevant for the envisioned task. Instead, we favor news language, regardless of from which section each text came from.

3.3. Selecting and training the annotators

Our subtopic annotation was performed by 14 annotators, all computational linguists with experience in corpus annotation, including undergraduate and graduate students, as well as professors.

The annotators went through training sessions during three days in one-hour daily meetings. During this step, the annotators were introduced to the task, its basis, and its relevance for NLP (for motivational purposes), segmented some news texts (which were not from the CSTNews corpus, in order to avoid bias in the actual annotation) and discussed their annotations, concepts and definitions related to the task, as well as compared their annotations. These training sections were conducted by two experienced annotators (who already had contact with the matter and performed subtopic segmentation in previous occasions).

We believe that all the annotators had some intuitive notion about the task. The three days of training proved to be enough for the annotators to acquire maturity in the process and to achieve a satisfactory agreement. This agreement was empirically checked during the discussions. As in other studies, as expected, agreement among judges was not perfect, since there is a (healthy) degree of subjectivity in the task (once it involves reading and interpreting texts), but overall agreement might be clearly observed.

3.4. Specifying the annotation procedure

In this work, the annotation phase took seven days in a daily one-hour meeting basis. On every annotation day, the annotators were organized in two groups, with at least five or seven annotators. The groups were randomly formed in each annotation session in order to avoid bias in the process. The annotators were instructed to read each text and to split it into subtopic blocks. The annotation was individual and the participants might not dialogue with each other. This process is similar to Hearst's methodology (1997).

The segmentation granularity was not defined and a subtopic could be formed by one or more sentences. Paragraph boundaries should not be taken into account, since a subtopic may be described inside a paragraph with other subtopics or even in more than one paragraph. One of the purposes of this work was to verify the subtopic segmentation

in news texts to check how they are distributed and whether or not paragraph boundaries are relevant for this task.

For each subtopic, one annotator was asked to identify it with a brief description using keywords. The description should be inserted in a concise and representative tag such as <*t label=”keywords”*>. The boundaries identified by the majority of the annotators were assumed to be actual boundaries.

3.5. Design of the annotation interface

Hovy and Lavid (2010) say that a computer interface designed for annotation must favor speed at the same time that it avoids bias in the annotation. Leech (2005), in turn, says that we may annotate texts using a general-purpose text editor, but this means that the job has to be done by hand, which may cause it to be slow and prone to error.

Since there is not a large set of tags for our annotation (only one), we did not develop a specific interface. The annotators used their preferred text editor and added a tag whenever they found a topic boundary. Thus, the annotators had the ability to go back and look over the parts that they had already looked at and change markings if desired, because they were manipulating a tool they already knew.

On the other hand, allowing the annotators to use the editor they were more familiar with meant that we had to check the encoding (since different editors and operating systems may use varied encoding standards, as Unicode and UTF-8), to make it uniform.

3.6. Choosing and applying evaluation measures

The underlying premise of an annotation is that if people cannot agree enough, then either the theory is wrong (or badly stated or instantiated), or the process itself is flawed (Hovy and Lavid, 2010). At first, it may seem intuitive to determine possible subtopic boundaries, but the task is very subjective and levels of agreement among humans tend to be low (see, e.g., Hearst, 1997; Passonneau and Litman, 1997; Sitbon and Bellot, 2006; Kazantseva and Szpakowicz, 2012). Agreement also varies depending on the text genre/type that is segmented. For example, technical reports have headings and subheadings, while other genres, such as news texts, have little demarcation.

The quality of annotation may refer to the agreement and consistency with which it is applied. As adopted by Hearst (1997), we used the traditional kappa measure (Carletta, 1996), which is better than simple percent agreement measures because it subtracts from the counts the expected chance agreement among judges. The kappa measure produces results up to 1, when agreement is perfect. It is assumed in the area that a 0.60 value is enough to the annotation to be reliable and conclusions may be drawn; however, it is known that such value highly depends on the subjectivity of the task at hand. In our case, we expect a lower value, since determining subtopics boundaries is a difficult task.

From left to right, Table 1 shows the days of annotation, the groups of annotators (represented by A and B letters), the number of annotators in each group, the number of texts that were annotated in each group per day, and the obtained agreement value. For instance, we see that the first day produced the best agreement among annotators, with a

0.656 agreement value for group A and 0.566 for group B. On the other hand, the lowest agreement was in the second day, with 0.458 for group A and 0.447 for group B. It is difficult to explain the fluctuations according to the day, but it may be the case that, on the first day, the annotators benefitted from the recent training, and that, on the second day, their confidence faltered as they encountered new cases. Agreement measures after the second day seem to improve, with some fluctuation. The average agreement was 0.560. Since this task is very subjective, such agreement is considered satisfactory. However, it is a bit lower than the agreement score obtained by Hearst (1997) in her seminal work, which was 0.647. This may be explained by the fact that her work used fewer texts (only 12), which were expository texts, where subtopic boundaries are usually more clearly indicated.

Table 1. Agreement values

Day	Groups	Number of annotators	Texts per group	Kappa
1	A	6	10	0.656
	B	7		0.566
2	A	5	10	0.458
	B	5		0.447
3	A	7	10	0.515
	B	5		0.638
4	A	5	10	0.544
	B	7		0.562
5	A	5	10	0.643
	B	5		0.528
6	A	5	12	0.570
	B	5	13	0.549
7	A	5	15	0.611
			Average	**0.560**

From the annotated texts, the reference segmentation was established. We computed the opinion of the majority (half plus one) in the boundaries. We adopted this strategy because we were looking for more course-grained subtopic boundaries, and we believed that these would have the consensus of most of the annotators. It is interesting that we observed only two cases with total agreement and others with more or less variation on segmentation. The variation in segmentation is related to factors such as the interpretation of the text and prior knowledge about the subject, mainly.

As an example of segmentation, Figure 2 shows different segmentations for the text in Figure 1. The rows numbered from 1 to 5 represent the segmentation made by each of the five annotators. Each box represents a sentence and the segmentation is indicated by vertical lines. The last line, labeled “Final”, represents the reference segmentation, obtained from the majority of the annotators.

One may see, for instance, that the first annotator did not place any subtopic boundary. The last boundary, after the last sentence, is expected, since the text ends. The second annotator placed boundaries after the fifth and sixth sentences, besides the one in the end of the text. These boundaries were the most indicated by the annotators and, therefore, were considered the ideal segmentation for the text, as shown in the last row.

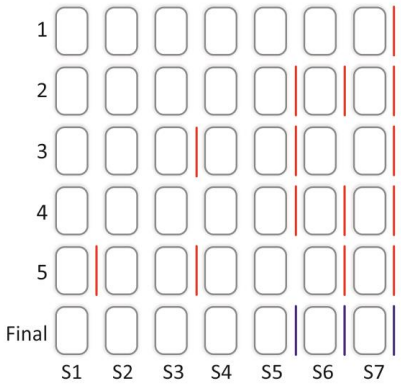

Figure 2. Example of different segmentations

Figure 3 shows the number of subtopics in the reference segmentation. It may be seen that there were eight texts (6% of the corpus) with only one subtopic, 24 texts (17%) with 2 subtopics, 50 texts (36%) with 3 subtopics, 32 texts (23%) with 4 subtopics, 18 texts (13%) with 5 subtopics, 4 texts (3%) with 6 subtopics, 2 texts (1%) with 7 subtopics, and 2 texts (1%) with 8 subtopics. Overall, the average number of subtopic boundaries in a text is 3. Most of them (99%) happen among paragraphs. This is related to how people structure their writings, using paragraphs as basic discourse units for the organization of the text. The descriptions given by each judge after a topic boundary were not used to define the final annotation. In this study, the descriptions were used only for better understanding the annotators' decisions.

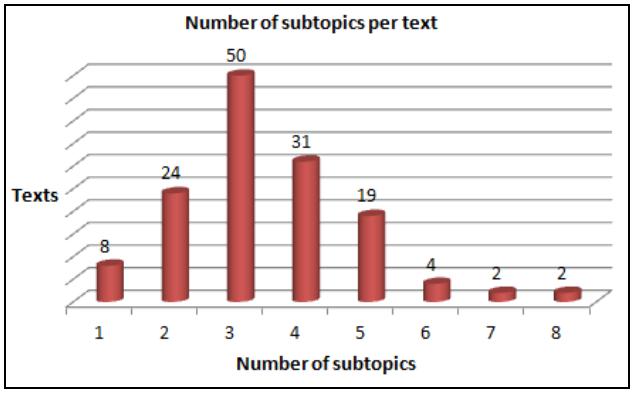

Figure 3. Number of subtopics in the texts

3.7. Delivering and maintaining the product

The corpus and its annotation are made available for research purposes. For each text, we provide the reference annotation and all the segmentations performed by each annotator. We considered important to make all these data available for researchers that are interested in investigating not only subtopic segmentation, but the influence of other text characteristics on each annotator behavior.

The annotation data is available in plain text format, which we adopted due to its simplicity. Even though the task of subtopic segmentation is very subjective, its representation (once it is done) is very straightforward.

4. Final remarks

This paper presented the main questions regarding corpus annotation for the phenomenon of subtopic segmentation. Our main contributions are two-fold: discussing and performing the annotation process in a systematic way, and making available a valuable reference corpus for subtopic study.

The corpus does not only contain the reported annotations, but several other annotations and information that are useful for several NLP tasks. It also includes single and multi-document discourse annotation, text-summary alignments, different types of summaries for each text and cluster, temporal and aspect annotations, and word sense annotation for nouns, among others. In the future, it is possible to look for annotation correlations, e.g., the correspondence of subtopic changes with discourse organization.

There are certainly several other research questions that deserve some attention in the future. For instance, it may be interesting to investigate whether annotators with different backgrounds identified different subtopic boundaries, or what the subtopic boundaries that were not identified by the majority of annotators may be indicating.

Acknowledgments

The authors are grateful to FAPESP, CAPES, CNPq and Natural Sciences and Engineering Research Council of Canada (Discovery Grant 261104-2008) for supporting this work.

References

Burger, S.; MacLaren, V.; Yu, H. (2002) The ISL meeting corpus: The impact of meeting type on speech style. In: *Proceedings of the International Conference Spoken Language Processing*, pp. 1-4.

Cardoso, P.C.F.; Maziero, E.G.; Castro Jorge, M.L; Seno, E.M.R.; Di Fellipo, A.; Rino, L.H.M.; Nunes, M.G.V.; Pardo, T.A.S. (2011) CSTNews – A discourse-annotated corpus for single and multidocument summarization of texts in Brazilian Portuguese. In: *Proceedings of the 3rd RST Brazilian Meeting*, pp. 88-105.

Carletta, J. (1996) Assessing Agreement on Classification Tasks: the Kappa Statistic. *Computational Linguistics*, Vol. 22, N. 2, pp. 249-254.

Carlson, L. and Marcu, D. (2001) *Discourse Tagging Reference Manual*. Technical Report ISI-TR-545. University of Southern, California.

Chang, T-H and Lee, C-H. (2003) Topic segmentation for short texts. In: *Proceedings of the 17th Pacific Asia Conference Language*, pp. 159-165.

Choi, F.Y.Y. (2000) Advances in domain independent linear text segmentation. In: *Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics,* pp. 26-36.

Galley, M.; Mckeown, K.; Fosler-Lussier, E.; Jing, Hongyan. (2003) Discourse segmentation of multi-party conversation. In: *Proceedings of the 41st Annual Conference of the Association for Computational Linguistics*, pp. 562-569.

Gruenstein, A.; Niekrasz, J.; Purver, M. (2007) Meeting structure annotation: Annotations collected with a general purpose toolkit. In: *Recent Trends in Discourse and Dialogue*. Series Text, Speech and Language Technology Springer Dordrecht, Vol. 39, pp. 247-274.

Hearst, M. (1997) TextTiling: Segmenting Text into Multi-Paragraph Subtopic Passages. *Computational Linguistics*, Vol. 23, N. 1, pp. 33-64.

Hennig, L. (2009) Topic-based multi-document summarization with probabilistic latent semantic analysis. In: *Recent Advances in Natural Language Processing*, pp. 144-149.

Hovy, E. (2009) Text Summarization. In: Ruslan Mitkov. *The Oxford Handbook of Computational Linguistics*, pp. 583-598. United States: Oxford University.

Hovy, E. and David, J. (2010) Towards a science of corpus annotation: a new methodological challenge for Corpus Linguistics. *International Journal of Translation Studies*, Vol. 22, N. 1, pp. 13-36.

Janin, A.; Baron, D.; Edwards, D.E.; Gelbart, D.; Morgan, N.; Peskin, B.; Pfau, T.; Shriberg, E.; Stolcke, A.; Wooters, C. (2003) The ICSI meeting corpus. In: *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, pp. 364-367.

Kazantseva, A. and Szpakowicz, S. (2012) Topical Segmentation: a study of human performance and a new measure of quality. In*: Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 211-220.

Koch, I.G.V. (2009) *Introdução à linguística textual*. São Paulo: Contexto.

Leech, G. (2005) Adding linguistic annotation. In: Martin Wynne. *Developing Linguistic Corpora: a guide to good practice*, pp. 25-38. Oxford: Oxbow Books.

Mohri, M.; Moreno, P.; Weinstein, E. (2010) Discriminative topic segmentation of text and speech. *Journal of Machine Learning*, Vol. 9, pp. 533-540.

Oh, H-J; Myaeng, S.H.; Jang, M-G. (2007) Semantic passage on sentence topics for question answering. *Information Sciences*, Vol. 177, N. 18, pp. 3696-3717.

Passonneau, R.J. and Litman, D.J. (1997) Discourse segmentation by human and automated means. *Computational Linguistics*, Vol. 23, N. 1, pp. 103-109.

Prince, V and Labadié, A. (2007) Text segmentation based on document understanding for information retrieval. In: *Proceedings of the 12th International Conference on Applications of Natural Language to Information Systems*, pp. 295-304.

Riedl, M. and Biemann, C. (2012) TopicTiling: a text segmentation algorithm based on LDA. In: *Proceedings of the 2012 Student Research Workshop*, pp. 37-42.

Sitbon, L. and Bellot, P. (2006) Tools and methods for objective or contextual evaluation of topic segmentation. In: *Proceedings of the 5th Edition of the International Conference on Language Resources and Evaluation*, pp. 2498-2503.

Wan, X. (2008) An exploration of document impact on graph-based multi-document summarization. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 755-762.