emeraldinsight

International Journal of Intelligent Computing and Cybernetics

Exogenous approach to improve topic segmentation

Marwa Naili, Anja Habacha Chaibi, Henda Hajjami Ben Ghezala,

Exogenous approach to improve topic segmentation

Marwa Naili, Anja Habacha Chaibi and
Henda Hajjami Ben Ghezala
*RIADI Laboratory, National School of Computer Science,
University of Manouba, Tunisia*

Received 5 January 2016
Revised 22 April 2016
Accepted 28 May 2016

Abstract

Purpose – Topic segmentation is one of the active research fields in natural language processing. Also, many topic segmenters have been proposed. However, the current challenge of researchers is the improvement of these segmenters by using external resources. Therefore, the purpose of this paper is to integrate study and evaluate a new external semantic resource in topic segmentation.

Design/methodology/approach – New topic segmenters (TSS-Onto and TSB-Onto) are proposed based on the two well-known segmenters C99 and TextTiling. The proposed segmenters integrate semantic knowledge to the segmentation process by using a domain ontology as an external resource. Subsequently, an evaluation is made to study the effect of this resource on the quality of topic segmentation along with a comparative study with related works.

Findings – Based on this study, the authors showed that adding semantic knowledge, which is extracted from a domain ontology, improves the quality of topic segmentation. Moreover, TSS-Ont outperforms TSB-Ont in terms of quality of topic segmentation.

Research limitations/implications – The main limitation of this study is the used test corpus for the evaluation which is not a benchmark. However, we used a collection of scientific papers from well-known digital libraries (ArXiv and ACM).

Practical implications – The proposed topic segmenters can be useful in different NLP applications such as information retrieval and text summarizing.

Originality/value – The primary original contribution of this paper is the improvement of topic segmentation based on semantic knowledge. This knowledge is extracted from an ontological external resource.

Keywords Semantics, Domain ontology, External resources, Topic segmentation

Paper type Research paper

1. Introduction

Topic segmentation is one of the most important tasks in natural language processing. It allows the partition of textual documents into segments, in a way that each segment deals with a specific topic. Regarding the importance of the topic segmentation, many topic segmenters have been presented. Almost all of these segmenters follow an endogenous approach that exploits only the text content, such as lexical repetition. However, this approach has proved its limits: it consists on the absence of external knowledge. To address this problem, some external resources were used to enhance the topic segmentation such as co-occurrence networks, thesaurus and semantic spaces. Even though, we notice that the choice of these external resources is very restricted. The major purpose of this paper is to exploit a new external resource in topic segmentation: a domain ontology, a resource which intends to add a semantic level to the topic segmentation.

This paper is organized as follows: Section 2 presents the literature review about topic segmentation; Section 3 explains the choice of the external resource; Section 4 focusses on the proposed topic segmenters TSS-Ont and TSB-Ont; Section 5 describes the ontology of test. Section 6 deals with the evaluation of these segmenters; Section 7 is dedicated to the conclusion and future works.

International Journal of Intelligent Computing and Cybernetics
Vol. 9 No. 2, 2016
pp. 165-178
© Emerald Group Publishing Limited
1756-378X
DOI 10.1108/IJICC-01-2016-0001

2. Related works

During the latest years, many English topic segmenters have been presented. As an example we mention TextTiling (Hearst, 1997), C99 (Choi, 2000) and F06 (Ferret, 2009). So far, all these segmenters are based on internal resources. Therefore, to enhance the quality of the topic segmentation, some researchers used external resources. We note that these external resources have been employed in two different ways. On the one hand, external resources have been used to improve existing topic segmenters, the work of Choi *et al.* (2001) can be mentioned as an example. In fact, they are the first ones to add external resource to their topic segmenter, named C99 (Choi, 2000), by using the latent semantic analysis (LSA). LSA is a powerful statistical method for extraction semantic information. After being patented in 1988, it has been first used in information retrieval field (Deerwester *et al.*, 1990). In topic segmentation, Choi *et al.* (2001) have proved that LSA improves the quality of their segmenter. Moreover, LSA is also used by Bestgen (2006) to reanalyze the same work as Choi *et al.* (2001). To go further, Bestgen and Pierard (2006) used LSA for another segmenter: TextTiling (Hearst, 1997). As result, they proved that, by adding semantic knowledge, the quality of topic segmentation is improved especially for the identification of topic boundaries among concatenated documents. Another example is the segmenter of Ferret (2009). He used a lexical co-occurrence network as an external resource in his segmenters F06C and F06CT. These segmenters are, respectively, based on the endogenous topic segmenters F06 and F06T. F06 is based on word recurrence. Yet, F06T takes into account the topics of documents to be segmented. Consequently, Ferret (2009) proved that the difference between F06 and F06C is not significant. Though, for F06CT, the improvement is important.

On the other hand, external resources have been used in order to offer a new topic segmenter without relying on another existing segmenters. As example, we cite Brants *et al.* (2002) who used the probabilistic latent semantic analysis (PLSA) in their topic segmenter TopSeg. PLSA is actually inspired from the LSA method, and it was suggested by Hofmann (1999). PLSA is a statistical method based on latent class model. Accordingly, Brants *et al.* (2002) found that using PLSA enhances the quality of topic segmentation. In 2004, Stokes *et al.* (2004) proposed a new segmenter, SeLeCT, based on WordNet thesaurus as an external resource. In fact, they used it to add semantic links among words, which led to an improvement of the quality of the topic segmentation. Similarly, Labadie and Prince (2008) used Larousse thesaurus in his segmenter Transeg to add semantic knowledge to the segmentation process. As a result, they (Labadie and Prince, 2008) proved the performance of his segmenter. In 2009, Misra *et al.* (2009) proposed a new segmenter based on the latent dirichlet allocation (LDA) method. This method was presented by Blei *et al.* (2003) and it allows the documents to be presented as a mixture of several topics. Thus, by using LDA, Misra *et al.* (2009) developed the quality of the topic segmentation. Recently, Bayomi *et al.* (2015) used ontological similarity in the process of topic segmentation based on DBpedia ontology. The main idea of this work is to add a conceptual level to the topic segmentation along with the lexical level. In fact, Bayomi *et al.* (2015) employed an edge-counting approach to calculate similarity between two concepts which corresponds to the path length connecting these concepts. The results showed that using ontology is promising. More recently, Naili *et al.* (2016) reused the LSA method in topic segmentation.

The main goal of this work is to study the different parameters of LSA and their impact in this field. As result, they proved the high performance of LSA especially with the right choice of parameters Table I.

3. Choice of the ontology as an external resource

Ontology was defined for the first time by Gruber (1993): an ontology is an explicit specification of a conceptualization. In other words, an ontology is an explicit specification of a shared conceptualization that provides terminology and a formal description of the semantics of a domain-related knowledge. Thus, given its semantic contribution, ontologies (Uschold and King, 1996) are widely used in different fields. De Almeida Falbo *et al.* (2005), for example, used ontologies in software engineering environment. The idea is to add semantic knowledge to improve the assistance given by these environments in order to support the development and the maintenance of software. Another example is the work of Coustaty *et al.* (2011). In fact, they integrate ontologies in image processing algorithms in order to detect their semantics. Moreover, Sriharee (2014) used ontology to propose an auto-tagging articles methodology. The tagging is based on an ontology in order to use articles' meaning. Also, Nebhi (2012) used ontology to annotate French newspapers articles. Actually, he performed an automatic semantic annotation based on semantic knowledge, which is deduced from the DBpedia ontology. Moreover, Bayomi *et al.* (2015) used The DBpedia ontology in topic segmentation to add a conceptual level. Recently, clinical data ontologies have been used by Gebremeskel *et al.* (2016), as a data mining technique, in healthcare services, in particular patient safety care. The common point among these works is the successful use of ontologies in order to integrate a semantic level in these different fields. Therefore, in this work, we integrate a domain ontology in the topic segmentation in order to add external semantic knowledge to the segmentation process.

4. Proposed topic segmenters

In this paper, we propose two English exogenous topic segmenters named topic segmenter based on sentences with Ontology (TSS-Ont) and topic segmenter based on Blocs with Ontology (TSB-Ont). These segmenters are based, respectively, on the two well-known segmenters C99 (Choi, 2000) and TextTiling (Hearst, 1997) which are the most used segmenters on topic segmentation (Bestgen, 2006; Bestgen and Pierard, 2006;

Segmenters	Original segmenters	External resources
CWM (Choi *et al.*, 2001)	C99 (Choi, 2000)	LSA
TopSeg (Brants *et al.*, 2002)	–	PLSA
SeLeCT (Stokes *et al.*, 2004)	–	Wordnet thesaurus
C99-LSA (Bestgen and Pierard, 2006)	C99 (Choi, 2000)	LSA
TextTiling-LSA (Bestgen and Pierard, 2006)	TextTiling (Hearst, 1997)	LSA
Transeg (Labadie and Prince, 2008)	–	Larousse thesaurus
F06C (Ferret, 2009)	F06 (Ferret, 2009)	Co-occurrence networks
F06CT (Ferret, 2009)	F06T (Ferret, 2009)	Co-occurrence networks
Segmenter of Misra *et al.* (2009)	–	LDA
OntoSeg (Bayomi *et al.*, 2015)	–	DBpedia ontology
ToSe-LSA (Naili *et al.*, 2016)	–	LSA

Table I. English topic segmenters with external resource

Labadie and Prince, 2008; Ferret, 2009). In fact, these two endogenous segmenters are based on lexical cohesion. So, based on this and as shown in Figure 1, TSS-Ont and TSB-Ont go through three different steps: pre-processing, stem-processing based on domain ontology and segmentation:

(1) Pre-processing: the pre-processing step is a common step for the two topic segmenters TSS-Ont and TSB-Ont that allows the extraction of words, the elimination of stop words and the stemming process by using Porter's (1980) stemmer.

(2) Stem-processing based on a domain ontology: in this step, the domain ontology is used in the same way for the two topic segmenters TSS-Ont and TSB-Ont. In fact, for each stem, we look into whether it belongs to the domain ontology or not. So, if it does, it will be replaced by its parent class. If not, the stem remains intact. To better understand this step, an example is presented in Figure 2. In this example, we used two sentences as an input. "Cryptanalysis is used to break cryptographic security systems. That can be done even if the

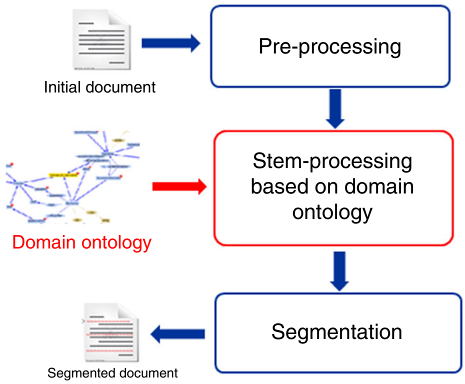

Figure 1.
Topic segmentation process based on a domain ontology

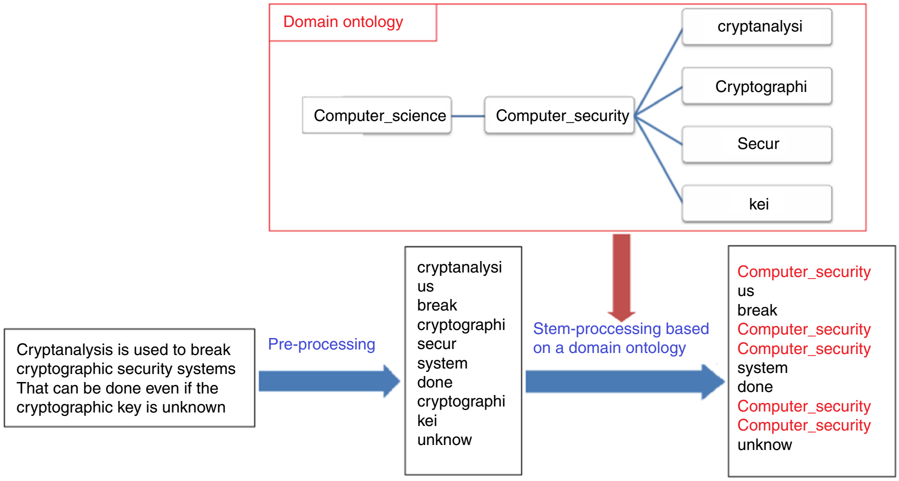

Figure 2.
Pre-processing and stem-processing based on a domain ontology steps with example

cryptographic key is unknown." So we conducted the first step, which is the pre-processing step. As result, stop words are eliminated, and the rest of the words are replaced by their stem according to Porter's (1980) stemmer. As an output of the pre-processing step, we have the following stems: cryptanalysi, us, break, cryptographi, secur, system, done, cryptographi, kei and unknow. This output is used as an input for the second step: the stem-processing based on a domain ontology. At this stage, we look if any of the obtained stems belong to the domain ontology. Consequently, we found out that cryptographi, cryptanalysi, secur and kei belong to the ontology and have computer security as a parent class. Therefore, only these four stems are replaced by this class. Thus, we notice that the contribution of using a domain ontology in topic segmentation lies in the identification of semantic links between words. Without using an ontology as an external resource, the four stems cryptographi, cryptanalysi, secur and kei are considered different words, and there is nolink between them. But, by using a domain ontology, a semantic link between these words is detected, and the similarity between these words will be increased.

(3) Segmentation: the third step of TSS-Ont and TSB-Ont is the segmentation process. This step is different for each segmenter. In fact, for TSS-Ont, the segmentation process is based on C99 (Choi, 2000) and it goes through four phases. The first one is the construction of the frequency dictionary. In this phase, each sentence is presented by a vector that is composed of stems and their frequencies. The second one is the construction of the similarity matrix. In this phase, the cosine measure is used to calculate similarity between sentences (Equation (1)). The third one is the construction of the rank matrix. Finally the last phase is the topic boundaries identification by using Reynar's maximization algorithm (Reynar, 1998). This algorithm detects boundaries by identifying the highest distribution of density:

$$\mathrm{Sim}\,(x,\,y) = \frac{\sum f_{x,j} f_{y,j}}{\sqrt{\sum f_{x,j}^2 \sum f_{y,j}^2}} \tag{1}$$

where $f_{x,j}$ (respectively, $f_{y},_{j}$) denote the frequency of word j in sentence x (respectively, y).

For TSB-Ont, the segmentation process is based on TextTiling (Hearst, 1997) and it goes through three phases. The first one is the bloc construction using a sliding window. The second one is the similarity calculation by using the cosine measure (Equation (1)). At this point, the similarity between blocs is calculated, and we note that a bloc is a set of sentences. Finally, the third one is the topic boundaries identification by using the cohesion score.

As a conclusion, we proposed two topic segmenters based, respectively, on C99 (Choi, 2000) and TextTiling (Hearst, 1997). We found out that the original segmenters use only a lexical level on topic segmentation. Yet, for the proposed segmenters TSS-Ont and TSB-Ont, we added a semantic level to the topic segmentation by taking into account the sematic links between stems. We also noticed that C99 and TextTiling take as input a raw text to segment. Yet, for TSS-Ont and TSB-Ont, a domain ontology is used as a second input with the raw text to segment. In the following section, we provide more detail about the used domain ontology.

5. Ontology for test

To propose an ontology for the computer science domain, we combined the two classifications of ArXiv[1] and ACM digital library[2]. In reality, these classifications contain almost all computer science disciplines. Therefore, as shown in Figure 3, we create an ontology which consists of 28 classes. These classes correspond to computer science disciplines such as artificial intelligence, information retrieval, databases and computer security. We observed that some of these 28 classes are composed of subclasses. For instance, we mention the databases class which contains five subclasses (management databases, data warehouse, data mining, data processing and security databases). The ground level of our ontology consists in stems representing the key words of the parent class. The stemming process is realized by Porter's (1980) stemmer. We note that, to cover the following relations: subsumption, composition and instantiation, we used these relations: Is-a, part-of and instance-of. We also note that this domain ontology is validated by four experts in the field from the RIADI laboratory[3].

6. Evaluation and discussion

To implement TSS-Ont and TSB-Ont, we opted to use java language. For the domain ontology construction, we used OWL on the Protege platform. To evaluate TSS-Ont and TSB-Ont, we compared these two letters with C99 and TextTiling. Therefore, we construct an English collection of 200 scientific papers. These papers are published in ArXiv and ACM digital library and dealt with the computer science field. With this corpus, we construct two different test sets: a set of documents and a set of segments. Each set contains 50 artificial documents. In fact, for the set of documents, an artificial document is a serial concatenation of two documents. However, for the set of segments, the concatenation is made in an alternate way by sections. The reason for using two different test sets is to increase the difficulty of the topic boundaries identification. It is really easier to detect boundaries between simple concatenated texts than between overlapping parts of texts. Thus, by using these two different test sets, we can study in depth the performance of the proposed segmenters. To report the result of this evaluation, we use the WindowDiff metric (Pevzner and Hearst, 2002) to measure the error rate.

6.1 TSS-Ont and C99

For the set of documents, the results of TSS-Ont and C99 are demonstrated in Figure 4 and Table II. Figure 4 describes the WindowDiff curves of these two topic segmenters. As a first statement, it is clear that the smallest error rates are obtained by TSS-Ont, except for few points. So, we can conclude that TSS-Ont outperforms C99. This result is confirmed by Table II, which presents the average results of WindowDiff. As shown in the same table, TSS-Ont is better than ArabC99 since it has the smallest value of WindowDiff (18.80 percent).

For the set of segments, the WindowDiff curves of TSS-Ont and C99 are shown in Figure 5. First of all, we can notice that these two curves are very close to each other. We can also remark that the biggest error rates are obtained by C99, except for some points. Table III, which contains the average results of WindowDiff, shows also that TSS-Ont has the smallest error rate (28.10 percent). We conclude that for the set of segments, TSS-Ont outperforms C99.

6.2 TSB-Ont and TextTiling

As shown in Figure 6 and Table IV, TSB-Ont outperforms TextTiling for the set of documents. In fact, in Figure 6, which shows the WindowDiff curves of these two segmenters, the highest error rates are obtained by TextTiling, except for few

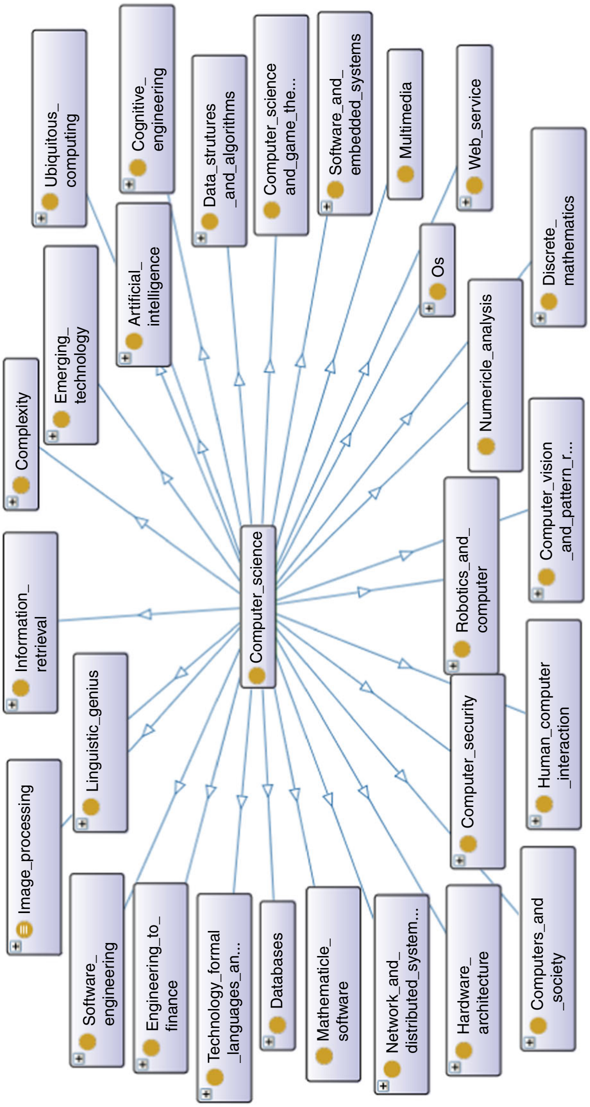

Figure 3.
The first level hierarchy of the used ontology

points. Besides, the Table IV, which contains the average results of WindowDiff, shows that TSB-Ont outperforms TextTiling by having the smallest value of WindowDiff (52.30 percent).

In the set of segments, we have the same result as in the first one. The Figure 7 describes the WindowDiff curves of TSB-Ont and TextTiling. We can notice that

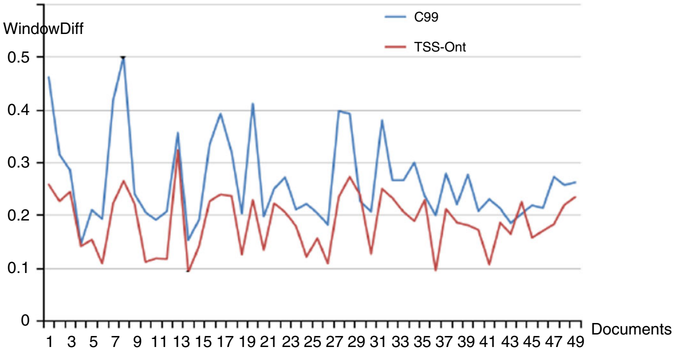

Figure 4.
WindowDiff results of TSS-Ont and C99 for the set of documents

Table II.
TSS-Ont and C99 results for the set of documents

Segmenters	WindowDiff (%)
C99	26.50
TSS-Ont	18.80

Note: WindowDiff results of TSS-Ont and C99 for the set of documents

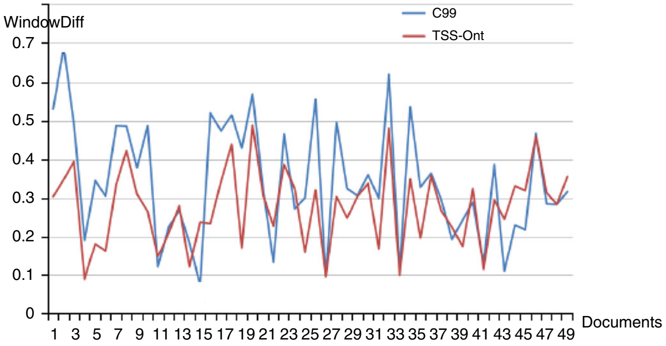

Figure 5.
WindowDiff results of TSS-Ont and C99 for the set of segments

Table III.
TSS-Ont and C99 results for the set of segments

Segmenters	WindowDiff (%)
C99	34.50
TSS-Ont	28.10

Note: WindowDiff results of TSS-Ont and bC99 for the set of segments

TSB-Ont has almost the lowest error rates compared to TextTiling. This result is confirmed in Table V: the average results of WindowDiff are described. Hence it is clear that TextTiling has the highest value of WindowDiff. Therefore, we can conclude that, for the set of segments, TSB-Ont outperforms TextTiling.

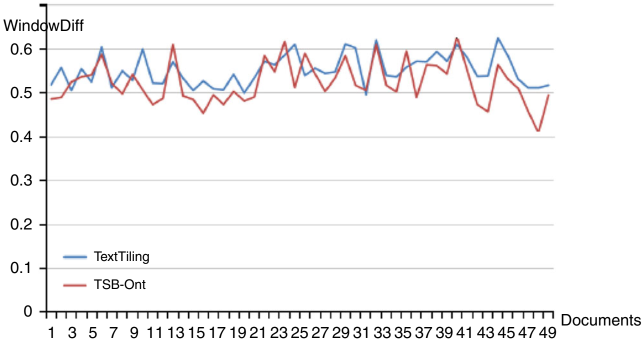

Figure 6. WindowDiff results of TSB-Ont and TextTiling for the set of documents

Table IV. TSB-Ont and TextTiling results for the set of documents

Segmenters	WindowDiff (%)
TextTiling	55.10
TSB-Ont	52.30

Note: WindowDiff results of TSB-Ont and TextTiling for the set of documents

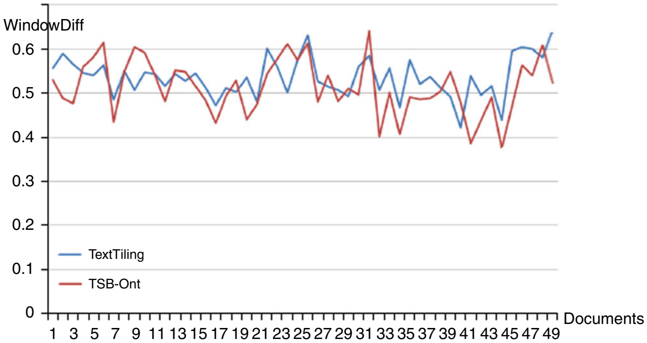

Figure 7. WindowDiff results of TSB-Ont and TextTiling for the set of segments

Table V. TSB-Ont and TextTiling results for the set of segments

Segmenters	WindowDiff (%)
TextTiling	53.60
TSB-Ont	51.40

Note: WindowDiff results of TSB-Ont and TextTiling for the set of segments

6.3 Discussion

As a first remark, we have proved that using domain ontology improves the topic segmentation. This improvement is detected by the high quality of topic segmentation. In fact, the experiment results of TSS-Ont and TSB-Ont show that the error rate decreases compared to C99 and TextTiling. This can be explained by the importance of semantic knowledge in topic segmentation. This knowledge takes into account the ontological similarity between words which are not detected with the lexical similarity. Yet, this improvement is not the same for TSS-Ont and TSB-Ont. Actually, using domain ontology reduces the error rate of TSS-Ont more than TSB-Ont. This can be explained by the fact that TSS-Ont computes the similarity between sentences that have a coherent meaning. So, by adding more semantic knowledge to these sentences, the detection of topics becomes easier. However, for TSB-Ont, the similarity is calculated between blocs. In this case, we can have several meanings in the same bloc. Therefore we can conclude that adding supplementary semantic knowledge improves topic segmentation especially if it is based on sentences like TSS-Ont. However, if it is based on blocs, the improvement is less important.

Moreover, it is interesting to note that, for both TSS-Ont and C99, the quality of the topic segmentation goes down for the set of segments compared to the set of documents. Yet, for both TSB-Ont and TextTiling, it is the opposite. But still, for TSB-Ont and TextTiling, the difference between the two test sets is not very significant. This can be explained by the fact that, like C99, TSS-Ont is more efficient when detecting segments with large size (the whole document) than smallest ones (just a section of a document). Like TextTiling, TSB-Ont does not depend on the size of segments.

On the other hand, we compared our work with Bayomi *et al.* (2015). In fact, best to our knowledge, they are the only ones who have used an ontological resource on topic segmentation. However, we note that these two comparisons are made on different corpus. The similarity between our works is the promising results of ontological resource on this field as shown in Table VI. Yet, the error rate obtained by TSS-Ont is smaller than OntoSeg which is the proposed segmenter of Bayomi *et al.* (2015). This can be explained by the fact that we used other relationships such as part-of and instance-of. Besides, computing the path length between two concepts may affect the relations of subsumption, composition, instantiation, etc. For TSB-Ont, the error rate of OntoSeg is better. The good performance of TSS-Ont and OntoSeg is explained by the fact that these two segmenters are based on sentences which is not the case for TSB-Ont.

To go further, we analyzed our results based on related works (Choi *et al.*, 2001; Bestgen and Pierard, 2006; Ferret, 2009). In these works, external resources have been used to improve existing topic segmenters. The evaluation resulting from these works are shown in Table VII. As a first remark, we noticed that only Bestgen and Pierard (2006) and ourselves used two test sets at the same time. Choi *et al.* (2001)

Table VI. Comparison between TSS-Ont, TSB-Ont and OntoSeg

Segmenters	WindowDiff (%)
OntoSeg	22
TSS-Ont	18.80
TSB-Ont	52.30

Source: Bayomi *et al.* (2015)

Segmenters	Improvement rate Documents (%)	Segments (%)	Corpus
CWM (Choi *et al.*, 2001)	33.33	–	Brown corpus 1993
C99-LSA (Bestgen and Pierard, 2006)	62.96	−2.17	Newspaper (Le Monde)
TextTiling-LSA (Bestgen and Pierard, 2006)	33.33	−2.08	Newspaper (Le Monde)
F06C (Ferret, 2009)	–	−0.33	CLEF 2003
F06CT (Ferret, 2009)	–	18.73	CLEF 2003
TSS-Onto	29.06	18.55	Scientific papers (ACM, ArXiv)
TSB-Onto	5.08	4.10	Scientific papers (ACM, ArXiv)

Table VII. Improvement rate results

showed that LSA improves the quality of topic segmentation by 33.33 percent for the set of documents. Yet, we note that they used the learning corpus as a test corpus in the same time, which may have distorted the results. Moreover, Bestgen and Pierard (2006) used LSA in their segmenters C99-LSA and TextTiling-LSA. They found that for the set of documents, LSA enhanced the quality of the topic segmentation in a remarkable way and especially for C99-LSA. Nevertheless, for the set of segments, the opposite occurred by having negative improvement rates. In fact, the good performance achieved by these segmenters for the set of documents is explained by the wide variety of topics (politics, archeology, philosophy, etc.) in the test set. Therefore, there was no such ambiguity to identify topic boundaries among concatenated documents dealing with completely different topics for the set of documents. Besides, Ferret (2009) evaluated F06C and F06CT only for the set of segments. As result, he has found that using a lexical co-occurrence network did not improve the topic segmentation for F06C. But for F06CT, a significant improvement is made by adding the topics of the document to segment to the co-occurrence network. However, we observed that for this segmenter, Ferret (2009) used a priori knowledge regarding the topic segmentation by using the topics of the documents. So this may threaten the validity of Ferret's (2009) results.

To synthesize these results, we can state that generally, using external resources improves the quality of topic segmentation. Besides, C99 is the most suitable segmenters for adding semantic knowledge. Furthermore, these segmenters are more efficient to detect the topic boundaries for the set of documents rather than the set of segments. Nonetheless, a question remains open: what is the best choice of external resource in topic segmentation? To answer this question, we need to study in depth these different segmenters by achieving a more reliable comparison with the same corpus.

7. Conclusion

In this paper, we used an ontological resource in topic segmentation in order to study its contribution and its effects. As a matter of fact, we proposed two topic segmenters TSS-Ont and TSB-Ont which are based, respectively, on C99 and TextTiling. TSS-Ont and TSB-Ont are based on a domain ontology as an external resource. For the evaluation, we used an English collection of scientific papers, and so we showed that using ontological resource improves the quality of topic segmentation especially for TSS-Ont. Thus, the main contribution of this work is the proposal of a new exogenous approach for topic segmentation. The novelty here lies in the successful integration of semantic knowledge based on a domain ontology. Furthermore, we are working on another approach of topic segmentation based on latent semantic methods (simple and probabilistic). The purpose of these different approaches is to achieve a more efficient comparison of different external resources in this field. In addition, we will automate the ontology construction and why not incorporate larger knowledge bases such as Wordnet, FreeBase, etc. We will also use a benchmark to evaluate and to carry out more significant experiments, as well as investigating the contribution of the ontological resource on Arabic topic segmentation by using two segmenters: ArabC99 and ArabTextTiling (Habacha *et al.*, 2014). And why not propose a multilingual hybrid topic segmenter.

Notes

1. CoRR Computing Research Repository. http://arxiv.org/corr/home
2. The ACM Computing Classification System (CCS). http://dl.acm.org/ccs.cfm?CFID= 617906392 & CFTOKEN=50956468
3. RIADI Laboratory. www.riadi.rnu.tn/

References

Bayomi, M., Levacher, K., Ghorab, M.R. and Lawless, S. (2015), "OntoSeg: a novel approach to text segmentation using ontological similarity", *IEEE International Conference on Data Mining Workshop, ICDMW 2015*, *Atlantic City, NJ*, pp. 1274-1283.

Bestgen, Y. (2006), "Improving text segmentation using latent semantic analysis: a reanalysis of Choi, Wiemer-Hastings and Moore", *Computational Linguistics*, Vol. 32 No. 3, pp. 5-12.

Bestgen, Y. and Pierard, S. (2006), "Comment evaluer les algorithmes de segmentation thematique? Essai de construction d'un mmateriel de reference", Actes de TALN: Verbum ex machina, Louvain-La-Neuve, Presse universitaire de Louvain, pp. 407-414.

Blei, D.M., Ng, A.Y. and Jordan, M.I. (2003), "Latent dirichlet allocation", *The Journal of Machine Learning Research*, Vol. 3, January, pp. 993-1022.

Brants, T., Chen, F. and Farahat, A. (2002), "Arabic document topic analysis", TREC, NIST, Gaithersburg, MD.

Choi, F.Y.Y. (2000), "Advances in domain independent linear text segmentation", *North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 26-33.

Choi, F.Y.Y., Wiemer-Hastings, P. and Moore, J. (2001), "Latent semantic analysis for text segmentation", *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 109-117.

Coustaty, M., Bouju, A., Bertet, K. and Louis, G. (2011), "Using ontologies to reduce the semantic gap between historians and image processing algorithms", *International Conference on Document Analysis and Recognition (ICDAR)*, IEEE, Beijing, pp. 156-160.

De Almeida Falbo, R., Ruy, F.B. and Dal Moro, R. (2005), "Using ontologies to add semantics to a software engineering environment", *17th International Conference on Software Engineering and Knowledge Engineering, Taipei*, pp. 151-156.

Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K. and Harshman, R. (1990), "Indexing by latent semantic analysis", *Journal of the American Society for Information Science*, Vol. 41 No. 6, pp. 391-407.

Ferret, O. (2009), "Improving text segmentation by combining endogenous and exogenous methods", *International Conference of Recent Advances in Natural Language Processing (RANLP), Borovets*, pp. 88-93.

Gebremeskel, G.B., Yi, C., He, Z. and Haile, D. (2016), "Combined data mining techniques based patient data outlier detection for healthcare safety", *International Journal of Intelligent Computing and Cybernetics*, Vol. 9 No. 1, pp. 42-68.

Gruber, T.R. (1993), "A translation approach to portable ontology specifications", *Knowledge Acquisition*, Vol. 5 No. 2, pp. 199-220.

Habacha, A.C., Naili, M. and Sammoud, S. (2014), "Topic segmentation for textual document written in Arabic language", *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems, KES-2014, Procedia Computer Science, Gdynia*, Vol. 35, pp. 437-446.

Hearst, M.A. (1997), "TextTiling: segmenting text into multi-paragraph subtopic passages", *Computational Linguistics*, Vol. 23 No. 1, pp. 33-64.

Hofmann, T. (1999), "Probabilistic latent semantic analysis", *Proceedings of Uncertainty in Artificial Intelligence, Stockholm*, pp. 289-296.

Labadie, A. and Prince, V. (2008), "Lexical and semantic methods in inner text topic segmentation: a comparison between C99 and Transeg", *NLDB*, pp. 347-349.

Misra, H., Yvon, F., Jose, J.M. and Cappe, O. (2009), "Text segmentation via topic modeling: an analytical study", *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, pp. 1553-1556.

Naili, M., Habacha, A.C. and Ben Ghezala, H.H. (2016), "Parameters driving effectiveness of LSA on topic segmentation", *17th International Conference on Intelligent Text Processing and Computational Linguistics CICLing, Springer LNCS Series, Lecture Notes in Computer Science, Konya, April 3-9*.

Nebhi, K. (2012), "Ontology-based information extraction for French newspaper articles", *35th German Conference on Artificial Intelligence, KI 2012: Advances in Artificial Intelligence, Springer Berlin, Heidelberg*, pp. 237-240.

Pevzner, L. and Hearst, M.A. (2002), "A critique and improvement of an evaluation metric for text segmentation", *Computational Linguistics*, Vol. 28 No. 1, pp. 19-36.

Porter, M.F. (1980), "An algorithm for suffix stripping", *Program*, pp. 130-137.

Reynar, J. (1998), "Topic segmentation: algorithms and application", PhD thesis, Computer and Information Science, University of Pennsylvania, Pennsylvania.

Sriharee, G. (2014), "An ontology-based approach to auto-tagging articles", *Vietnam Journal of Computer Science*, Vol. 2 No. 2, pp. 85-94.

Stokes, N., Carthy, J. and Smeaton, A.F. (2004), "SeLeCT: a lexical cohesion based news story segmentation system", *AI Communications*, Vol. 17 No. 1, pp. 3-12.

Uschold, M. and King, M. (1996), "Ontologies: principles, methods and applications", *Knowledge Engineering Review*, Vol. 11 No. 2, pp. 93-155.

About the authors

Marwa Naili received a Computer Science Engineering Diploma from the National School of Computer Science (ENSI), University of Manouba, Tunisia in 2013. Currently she is a PhD Student in RIADI Laboratory, ENSI since 2014. Her research interests are natural language processing (NLP), text mining and information retrieval. Marwa Naili is the corresponding author and can be contacted at: maroua.naili@riadi.rnu.tn

Anja Habacha Chaibi received in 1988 a Computer Science Engineering Diploma from the University of Tunis, in 1989 a Master's Degree from the Henri Poincare University (France), in 1993 a PhD Degree in Computer Science from the National Polytechnic Institute of Lorraine (France). She is an Assistant Professor at the National School of Computer Science, a Research Fellow of RIADI Laboratory and a Committee Member of TCPC. Her research interests are information retrieval, NLP and scientometrics.

Henda Hajjami Ben Ghezala is a Professor of Computer Science at the National School of Computer Sciences (ENSI), University of Manouba, Tunisia. She is the Head of the Research Laboratory RIADI (softwaReengIneering, distributed applications, decision systems and intelligent imaging). Her main research interests are intelligent information systems and software engineering. She has organized several workshops and conferences on computer science and software engineering.