Subtopic Segmentation of Scientific Texts: Parameter Optimisation

Natalia Avdeeva[1], Galina Artemova[2], Kirill Boyarsky[2], Natalia Gusarova[2(✉)], Natalia Dobrenko[2], and Eugeny Kanevsky[1]

[1] Saint Petersburg Institute for Economics and Mathematics, Russian Academy of Sciences, Saint Petersburg, Russia
[2] Saint Petersburg National Research University of Information Technologies, Mechanics and Optics (ITMO University), Saint Petersburg, Russia
natfed@list.ru

Abstract. Information research within a scientific text needs to deal with the problem of automatic document partition on subtopics by taking text specifics and user purposes into account. This task is important for primary source selection, for working with texts in foreign languages or for getting acquainted with research problems. This paper is focused on the application of subtopic segmentation algorithms to real-life scientific texts. For studying this we use monographs on the same subject written in three languages. The corpus includes several original and professionally trasnlated fragments. The research is based on the TextTiling algorithm that analyses how tightly adjoining parts of the text cohere. We examine how some parameters (the cutoff rate, the size of moving window and of the shift from one block to the next one) influence the segmentation quality and define the optimal combinations of these parameters for several languages. The studies on Russian suggest that external lexical resources notably improve the segmentation quality.

Keywords: Text tiling · Classification · Parsing · Segmentation

1 Introduction

The coverage of relevant information sources substantially predetermines the efficiency in research work, particularly in data intensive fields. The sources in general include scientific texts, such as monographs, textbooks, scientific articles etc. As a rule, all of them are large information-rich documents in the original with a typical structure [17,19].

All over the world, the scientists take measures to share primary scientific sources via the Internet. However, the efficiency of the information retrieval in this corpus is still of poor quality. Their structures cannot always be presented by search attributes, which are traditional for the Web (meta tags, keywords etc.). As a result, a user can get either a full document where he has to find information manually by himself or a detached extract with the greatest keyword frequency rate. In the latter case, it is hard to form a clear picture of a document topic.

P. Klinov and D. Mouromtsev (Eds.): KESW 2015, CCIS 518, pp. 3–15, 2015.
DOI: 10.1007/978-3-319-24543-0_1

Thus, it is necessary to organize information research within a scientific text. It needs in turn to solve a problem of automation document partition on subtopics taking into account text specifics and the purposes of the users. This task is important for primary source selection, for working with texts in foreign languages or for quick acquaintance with research problems.

Many approaches to topic segmentation of the text have already been described. You can see their brief review in Sec. 2. As a rule, they are quite effective with composed texts like concatenated separate sentences or short reports from newspapers or Internet sources ([3,5,10]) or with large text corpora ([4,8,23]). Meanwhile, results of applying these methods to real scientific texts are rather scantly and contradictory ([5,10,12]).

This paper is focused on application of topic segmentation algorithms to real scientific texts. For studying this we used monographs on the same subject written in three languages. The experimental set includes several fragments both in the original language and in professional translation. During our research, we varied lexical units for analysis, the cutoff rate, the size of moving window and of the shift from one block to the next one. We examined how these parameters, text language, inclusion/exclusion of external lexical resources (classifiers, stoplists etc.) influence the quality of segmentation.

2 Related Work

Almost all topic segmentation methods are based on text cohesion. According to work [25], the most dominated types of subtopic cohesion are lexical (repetition, synonymy and reference) and grammatical types (parallelism):

1. repetition is a usage of the same terms in adjacent sentences of the same subtopic;
2. synonymy refers to the doublets (terms that are close in meaning);
3. reference is the use of an expression (pronoun or a demonstrative pronoun) the meaning of which depends on the previous or next expression;
4. parallelism appears in revealing the thesis by sentences with parallel structure and the same form of their predicates.

Most of the existing methods are based on examining repetition. These methods can be divided in two groups. The ones of the first group use data of cohesion between adjacent parts. One of the most well-known techniques is TextTiling [12,13], that includes the following steps:

(a) text is lemmatized and stop words are removed. Hence the text is regarded as a sequence of N tokens;

(b) sequences of W tokens combine in pseudosentences. The k pseudosentences join in blocks, which then are used as sliding window with the step of s pseudosentences. In a standard technique $s = 1$. Hence, firstly a group of tokens from 0 to $(W \times k)$ is compared with the ones of tokens from W to $W \times (k+1)$. Then the latter group is compared with the ones of tokens from $(W \times 2)$ to $(W \times (k+2))$. It repeats until the second border is reached.

(c) lexical similarity of adjoining blocks is computed as the cosine of angle φ between $(W \times k)$-dimensional vectors:

$$\cos\varphi_i = \frac{\sum\limits_n w_{n,i-1} w_{n,i}}{\sqrt{\sum\limits_n w_{n,i-1}^2} \sqrt{\sum\limits_n w_{n,i}^2}}, \qquad 0 \le \cos\varphi \le 1 \tag{1}$$

where $w_{n,i}$ is a weight of n^{th} token in i^{th} block.

(d) local minima of (1) are regarded as the boundaries of the segments (rounded to the nearest sentence or paragraph).

The methods of the second group analyse the distribution of tokens repeated throughout the text. Thus, DotPlotting technique [22] examines text cohesion with the use of two-dimensional graph (named dotplot). The positions of tokens in the text are plotted on its X and Y. If a particular token appears in the positions x and y of the text then dots $(x, x), (y, y), (x, y)$ and (y, x) are plotted in the graph. At that, cohesive text segments visually correlate to squares with a high dot density along the diagonal. The resulting distribution is examined for extrema with the help of one of following strategies. You can either minimize dot density on the boundaries or maximize it within the segment. This idea was developed into C99 technique [4]. There the measure of lexical cohesion between tokens of adjacent segments is visualized in the same way and then maximum density areas of this measure are found by means of the dynamic programming.

A number of modifications to the standard techniques (see [5,7,9,15,20]) allows to analyze other types of cohesion besides repetition. For example, a type 2 can be found with the help of external vocabularies (WordNet) or by modifying the cosine measure with a coefficient reflecting word frequency in external document set (Internet) [5]. It is offered to combine the DotPlotting measures of lexical cohesion within the segment and between two segments to take type 4 into account [30].

Hence, now we can list obstacles to implementing discussed techniques into real search engines. First, most of these methods are developed for the English language. Topic segmentation task is less studied for other language groups [3,29]. Secondly, there are no open profound dictionaries of synonyms or other net resources for many languages including Russian. Thirdly, the efficiency of all these methods strongly depends on text cohesion argument, which is initially unknown.

The methods of text hierarchical segmentation have been developed in recent years [8,16,23]. Most of them are based on word cohesion model presented as a multidimensional word distribution by topics. At that, the occurrence of every word connects with one of several topics that are discussed in the text. The mathematical foundation of this approach is Latent Dirichlet Allocation (LDA) [1] widely used in the machine learning. For example, [8] presents a hierarchical Bayes algorithm revealing two levels of linear text segmentation. TopicTiling [23] as TextTiling is based on cosine cohesion measure of two adjacent segments.

However, it uses frequency of topic identifiers calculated for every word by LDA instead of word frequency.

The larger training set and the closer its statistical distributions to the text, the higher effectiveness these approaches show. Ideally, both training set and the analyzed text should be of the same domain [23]. It is hard to reach it while processing real scientific text: its value depends mainly on its uniqueness.

Here we examined which segmentation parameters are important for the purpose of developing topic segmentation of scientific texts. Our key method combines a linear segmentation and cosine cohesion measure between two segments.

3 Experiments

3.1 Text Selection and Pre-Processing

Texts were extracted from monographs on the same technical topics written in three languages (Russian, English and French). See Table 1 for details of sources.

Table 1. Features of text sources

Source label	Source bibliographical entry	Source language	Fragment's size (in printed chars)
1. Romme	Romme N. Ch. :L' Art de la Marine, u Principes t Préceptes Generaux d l'Art de Construire, d'Armer , de Manuvrer et de Conduire ds Vasseaux, par . Re/ ed. : P.-L. Chauvet. La Rochelle, 1787. Chapitres VII, VIII.	French	110261
2.Romme Rus	Romm N. Ch.: The Navy Art, or Principles and Basic Rules of the Shipbuilding, Equipment and Ship Handling by Romm/ A. Shishkov. Saint-Petersburg, 1793. (in Russian) Chapters 7, 8.	Russian	161152
3. U-boat	Williamson G., Johnson L.: U-Boat crew 1914-45/, ed. Osprey Publishing, Great Britain, 1995.	English	60563
4.U-boat Rus	Williamson G.: German Submarine Fleet. 19141945 / M.A. Maltseva, AST, Moscow, Russia, 2003. (in Russian)	Russian	60560
5. News	Concatenated news sources (the Internet)	Russian	23800

The corpus includes several fragments both in the original (lines 1, 3) and in professional translation (lines 2, 4). This approach solves a domain identification problem and allows studying language features of segmentation in pure form. Part-of-speech tagging of English and French texts was fulfilled by the means of net service OpenXerox[1]. The Russian texts were tagged using parser SemSin [14].

[1] Xeros Linguistic Tools: Part of Speech Tagging: http://xerox.bz/1HYXX1Q

Stop-lists for every text were formed manually with the help of frequency analysis. Besides we compiled 20 pieces of political news extracted from the Web in a random way (line 5 in Table 1). Sometimes we used a classifier [28] described in Sec. 4.5. See the methods of text pre-processing in Table 2.

Table 2. Pre-processing types

Abbreviation	Description
L	lemmatization
L+SL	lemmatization + stop words removing
L+N	lemmatization + POS tagging + noun selection
L+N+Adj+V	lemmatization + POS tagging + noun, adjective and verb selection
L+POS+Class	lemmatization + POS tagging + the use of external classifier [28]

3.2 Text Processing Method

We applied TestTiling as a basic processing technique because it allows to analyse cohesion of the texts and cosine similarity transparently. See above the processing steps and parameters abbreviation (p. 2). The next parameters were varied during the research: size of blocks [tokens] $W \times k$; overlapping size between blocks [tokens] $s \times k$. Besides, in some experiments we worked with blocks of variable length, which is equal to the paragraph length.

3.3 Segmentation Evaluation Metrics

A range of metrics (including precision-recall ratio [13], edit distance [21], $P\mu$ and Pk measure [6], WindowDiff) is proposed to estimate the quality of text segmentation. Each of them has specific limitations.

For example, WindowDiff compares the positions of segment boundaries, which were set according to the baseline, and of ones, determined by the algorithm, within a sliding window. Then the number of windows that were set as boundaries by mistake is normalized to a total number of windows. However, the subsequent studies [11,18,24] revealed limits of WindowDiff metrics. It equally evaluates false (false positive, *FP*) and missed (false negative, *FN*) boundaries, which should have different importance depending on the specific segmentation task. Besides, Window Diff ignores the rate of missed boundaries and emphasizes mistakes at the beginning and end of the text. Thus, modifications of Window Diff were proposed [24].

Here we used balanced *F*-score [2] to estimate the segmentations quality:

$$F = \frac{2 \times P \times R}{P + R} \tag{2}$$

where $P = \frac{TP}{H}$ is precision,$R = \frac{TP}{TP+FN}$ means recall, FP means a number of false boundaries, FN a number of missed boundaries and H a total number of found boundaries.

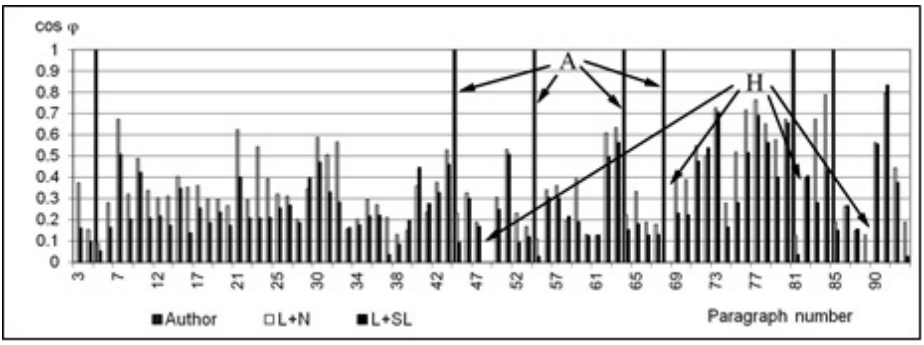

Fig. 1. Author segmentation (A) and automatically detected segment boundaries (H) of two pre-processing types for Romme

We used topic shift boundaries set by the author (A) as baselines (see Fig.1).

Selection of segmentation parameters carried out as follows. The cosine measure (1) was analyzed according to the cutoff rate z, which is on the ordinate axis (Fig. 1). The values of $\cos\varphi$ less than or equal to z, were considered as topic shift, i.e. a boundary between segments (H). Then the sequences A and H were compared. The local minima (valleys) for the same or adjacent segments of these sequences were labeled as true matching (TP). Such rounding is reasonable because, as shown by a detailed analysis, often in the first paragraph of a new topic the author tries to gradually change the subject, and in fact, the transition occurs only in the following paragraph. The other valleys H were marked as false ones (FP) while valleys A were labeled as missed ones. So our task was to determine the optimal cutoff rate z, on which we can easy divide a text into segments. To find the best value of F-score the cutoff rate z ranged from 0 to 1 with step of 0.05.

Note that the above metrics evaluate the segmentation quality only post factum and do not allow to fine-tune the parameters of the segmentation algorithm, while different tasks require different precision and recall ratio, which is especially important for scientific texts. This was one more reason for using F-score (2). It can be regarded as the optimization criterion with its maximum as the best P and R ratio. Moreover, it can be easily adjusted to the users needs by weighting P and R components.

4 Experimental Results and Discussion

4.1 The Influence of the Size of Overlaps Between Blocks

In standard TextTiling scheme [12] it is recommended to choose a sliding window of the size ($W \times k$) tokens, where k is a number of pseudosentences with the size of W tokens, and overlapping blocks $s = \frac{W \times k}{2}$. But according to our experiments on combining different values of s, W and k, any overlap lowers contrast range and valleys that indicate semantic boundaries practically disappear.

A typical example is shown in Fig. 2. On the plot the result of analyzing texts without overlaps is represented by the curve a. Due to the wide data spread there, semantic boundaries can be defined automatically. On the contrary, the curve b is to smooth to find the boundaries as it shows the result of analyzing texts with overlaps. Note that this result agrees with the conclusion reached in the pilot study [5]. Thus, we did not use overlaps in our researches.

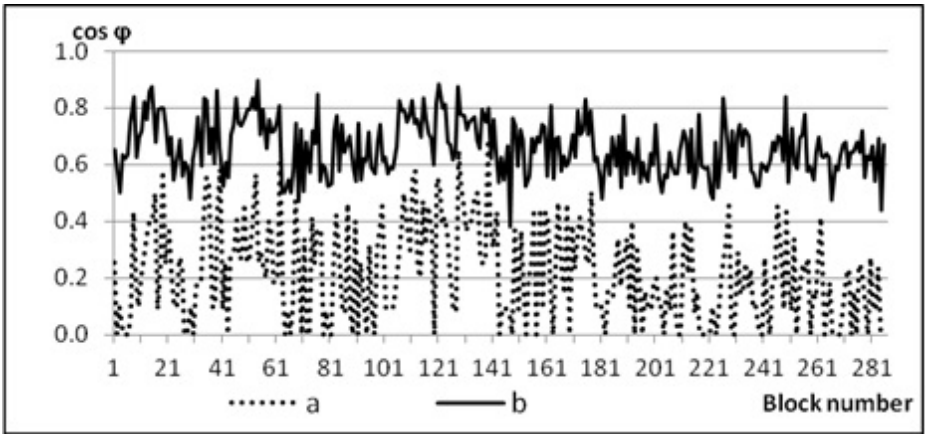

Fig. 2. Cosine measure for U-boat text with $L+N$ pre-processing type: the curve a: $(W \times k) = 10, s = 0$; the curve b: $(W \times k) = 20, s = \frac{W \times k}{2}$.

4.2 The Influence of Block Size

As noted above, the basic algorithm TextTiling [12,13] uses windows of fixed length as the analysis unit, while a resulting boundary position is extended to the nearest sentence or paragraph. Our experiments have shown that in this case the analysis quality does not meet the requirements of actual scientific text (see Table 3). In the case of short window size (10 nouns), word random changes lead to a large number of "false alarms" reducing the precision. However the larger a window size is (e.g. 40 nouns) the higher the probability is that the window overlaps a real boundary, resulting to recall decrease. See sample results for "U-boat" with $L+N$ pre-processing type in Table 3, var. 1-3.

Table 3. The influence of blocks size

Variants	1	2	3	4
Size of blocks [tokens]	10	25	40	paragraph
F-score	0.06	0.04	0.03	0.17

In our work, we examined the assumption that semantic borders should be placed at the boundaries of paragraphs. It should be noted that contradictory opinions on this matter could be found in the literature. For example, [27] suggests that: "... we can say that a paragraph in the scientific text is independent, graphically highlighted text unit containing a particular idea or its fragment". On the other hand, according to [26], the text division by formal components (paragraphs or sections) does not allow to identify subtopics. The boundaries of formal and semantic units may differ; paragraph divisions depend on the document type and purpose (e.g., text and art news). Large paragraphs may contain several subtopics.

We experimented with paragraph-sized blocks. Short paragraphs as a rule are equal to headings or lyrical digressions thus they were joined to the next one.

Analysis results show a dramatic increase of F-score in the comparison with ones on text with the fixed length of the segments (Table 3, var. 4).

Thus in our further experiments we divided all the texts by formal paragraphs.

4.3 The Influence of Cutoff Rate

In our researches, boundaries between segments were set in accordance with the cutoff rate z. The higher it gets the lower the precision P falls as missed boundaries appear; meanwhile the recall R increases as more and more true boundaries can be discovered. Thus, our task is to set the optimum cutoff rate in respect to F-score (Fig. 2). According to our experiments, the optimum cutoff rate z depends on the pre-processing type. Maximum F-score is achieved at $z = 0.1 \ldots 0.15$ for all types except *L+POS+Class* pre-processing (Table 2). In the latter case examined in Section 4.5 F-score gets its maximum at higher cutoff rate (Fig. 3).

We examined if pre-processing type correlates with the source language. The preprocessing type is obviously determines what words should be selected to analyse texts on different languages. Thus, it may depend on type of cohesion. In particular, we can find repetitions reference in *L+S* and *L+SL* (excluding the most frequent words). *L+N+Adj+V*, where the stop list was formed only from function words, indicates both repetitions and parallelism.

In given texts we excluded both standard stop words and the most frequent informative words: boat (лодка) and Germany (Германия) in Russian texts, U-boat, jacket, war (in English ones), voile (sail), poulie (block), mat (mast), fig (Figure), vergue (yard) in French texts.

See how different pre-processing types change the maximum F-scores for given texts in Table 4.

We can easily explain these results using charts of cosine cohesion measurements between adjoining paragraphs in "Romme" (Fig. 1).

Full vertical line graphs indicate the text division into subsections according to headings (author segmentation). According to the chart, the cosine measure is lower when stop words are excluded. Nevertheless, the "valleys" are on the

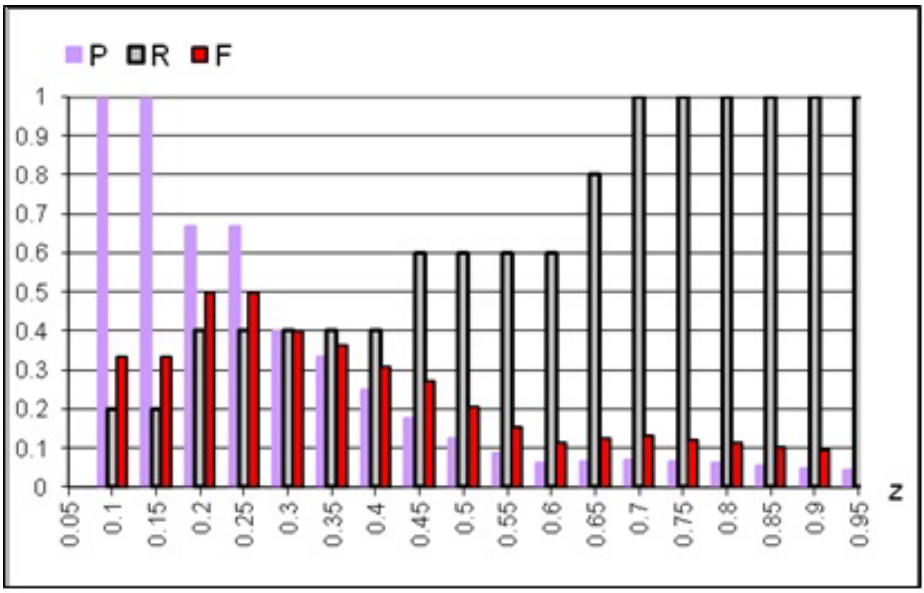

Fig. 3. The dependence of P, R and F-score on the cutoff rate z.

Table 4. F-scores for different pre-processing types

Source	*F-score*		
	L+N+Adj+V	*L+N*	*L+SL*
"U-boat Rus"	0.19	0.21	0.11
"U-boat"	0.24	0.17	0.19
"Romme"	0.46		0.53

same place. Thus, the segmentation quality does not change much. Compare the best F-scores for all nouns, including and excluding stop words: 0.46 at $z = 0.15$ and 0.53 at $z = 0.1$ correspondingly.

Thus, it is preferable to apply *L+N* type to Russian texts. In English texts, analysis of adjectives and verbs considerably improves the segmentation quality. In French texts analysis without stop words improves it a little.

See maximum F-scores of texts analyzed by words" (i.e. for all pre-processing options except *L+POS+Class*) at optimum cutoff rate in Table 5. Note that these values correlate with the ones in studies [3], [5].

Table 5. Maximum F-scores

Source	U-boat Rus	U-boat	Romme Rus	Romme	News
F-score	0.21	0.17	0.22	0.46	0.60

Table 5 shows that in scientific texts author segmentation on headings and subtitles does not correspond to vocabulary changes. The analyzer poorly detects them. The text is of a low contrast in respect to its vocabulary. On the other hand, the segmentation quality of news texts is considerably higher.

4.4 The Influence of External Lexical Resources

The stability (robustness) of segmentation with external lexical resources is known to increase. As external resources, we can use not only dictionaries of synonyms [5] but lexical databases [23]. There is no open dictionary of Russian synonyms of high quality, so we used a semantic classifier [28]. We suggested that words belonging to the same class would be marked as the same ones. The examples are vessel (судно), ship (корабль), frigate (фрегат) from "Romme Rus" and helmet (каска), service cap (фуражка), peakless cap (бескозырка) from U-boats Rus.

Let us consider a typical example and compare two sentences from adjoining parts of "U-boat Rus":

Ribbons of peakless caps for the Kaiserliche Marine had gilt and silver thread block lettering (for sea-going personnel and for administrative personnel respectively).

Ленточки бескозырок матросов кайзеровского флота имели надпись прописными печатными буквами, вышитыми золотой или серебряной канителью.

and

The field cap was cut from fine-quality navy blue cloth wool, usually with a black or dark blue cotton or artificial silk lining.

Пилотка кроилась из темно-синего плотного сукна, обычно с черной или темно-синей подкладкой из искусственного шелка.

These sentences in spite of belonging to the same subtopic have no lexical repetitions. Thus, its cosine measure computed "by words" is zero. However, words peakless cap (бескозырка), field cap (пилотка), lining (подкладка) are of the class Clothes and ribbon (ленточка), gild or silver thread (канитель), cloth wool (сукно), silk (шелк) belong to the class Fabrics. Hence, analyzing only nouns by classes we get $\cos\varphi = 0.71$.

Fig. 4 shows the functions of *F*-score from cutoff rate on the example of analyzing "Romme Rus" by words and by classes (the *L+POS+Class type*).

As one can see, *F*-score of pre-processing "by classes" increases more than twice. The accuracy level can range to reach the best result. For example, headdresses can be regarded as a separate class or a part of "Clothes". You can compare the best *F*-scores given in Table 6.

Thus, the results by classes are much higher than the one by words. It is notable that news texts are divided equally at $z = 0.5$ in any way. This means that $FN = FP$.

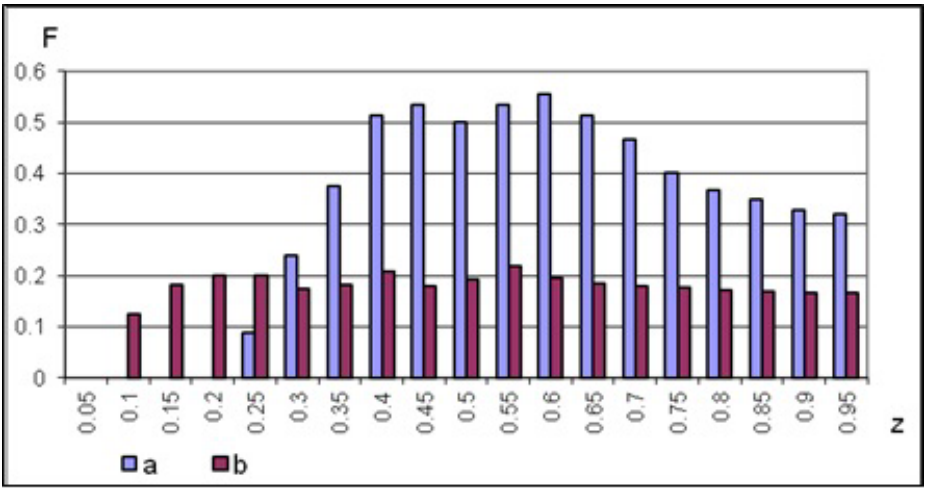

Fig. 4. The comparison of analysis "by classes" and "by words".

Table 6. *F*-score of Analysis "by words" and "by classes"

Text	U-boat Rus	Romme Rus	News
By words	0.21	0.22	0.60
By classes	0.50	0.7	0.78

Thus, to get the best results of automatic segmentation it is necessary to evaluate the similarities and differences between text fragments by packaged vocabulary and not by separate words. This allows to use all types of cohesion more efficient.

5 Conclusion

We studied the specifics of applying subtopic segmentation methods on real scientific texts on the same subject in three languages. The corpus includes several fragments both in the original language and in professional translation. The research is based on the TextTiling algorithm that analyses how tightly adjoining parts of a text cohere. We examined how some parameters (the cutoff rate, the size of moving window and of the shift from one block to the next one) influence the segmentation quality. The optimum combinations of these parameters are defined for several languages. The studies on Russian language suggest that external lexical resources notably improve the quality of segmentation.

References

1. Blei, D., Ng, A., Jordan, M.: Latent dirichlet allocation. Journal of Machine Learning Research **3**, 993–1022 (2003)
2. Bolshakova, E.I., Klyshinsky, E.S., Lande, D.V., Noskov, A.A., Peskova, O.V.: Automatic processing of natural language texts and computer linguistics. Moscow State Institute of Electronics and Mathematics (2011)
3. Chaibi, A., Naili, M., Sammoud, S.: Topic segmentation for textual document written in arabic language. In: Procedia Computer Science: 18th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems, KES2014, vol. 35, pp. 26–33 (2014)
4. Choi, F.: 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. 26–33 (2000)
5. Dias, G., Alves, E., Lopes, J.: Topic segmentation algorithms for text summarization and passage retrieval: an exhaustive evaluation. In: AAAI 2007 Proceedings of the 22nd National Conference on Artificial Intelligence, vol. 2, pp. 1334–1339 (2007)
6. Douglas, B., Berger, A., Lafferty, J.: Statistical models of text segmentation. Machine Learning **34**(1–3) (1999)
7. Du, L., Buntine, W., Johnson, M.: Topic segmentation with a structured topic model. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 190–200 (2013)
8. Eisenstein, J.: Hierarchical text segmentation from multi-scale lexical cohesion. In: NAACL 2009 Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 353–361 (2009)
9. Eisenstein, J., Barzilay, R.: Bayesian unsupervised topic segmentation. In: Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, pp. 334–343 (2008)
10. Flejter, D., Wieloch, K., Abramowicz, W.: Unsupervised methods of topical text segmentation for polish. In: Balto-Slavonic Natural Language Processing 2007, pp. 51–58 (2007)
11. Georgescu, M., Clark, A., Armstrong, S.: An analysis of quantitative aspects in the evaluation of thematic segmentation algorithms. In: SigDIAL 2006 Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue (2009)
12. Hearst, M.: Texttiling: Segmenting text into multi-paragraph subtopic passages. Computational Linguistics **23**(1), 33–64 (1997)
13. Hearst, M., Plaunt, C.: Subtopic structuring for full-length document access. In: SIGIR 1993: Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 59–68 (1993)
14. Kanevsky, E.A., Boyarsky, K.: Semantics and sintactics parser semsin. In: Dialog-2012: International Conference on Computational Linguistics (2012). http://www.dialog-21.ru/digests/dialog2012/materials/pdf/Kanevsky.pdf (date of access June 29, 2015)
15. Kazantseva, A., Szpakowicz, S.: Linear text segmentation using affinity propagation. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pp. 284–293 (2011)

16. Kazantseva, A., Szpakowicz, S.: Hierarchical topical segmentation with affinity propagation. In: Proceedings of COLING 2014, The 25th International Conference on Computational Linguistics: Technical Papers, pp. 37–47 (2014)
17. Kotyurova, M.P.: Scietific style of speech. Akademiya (2010)
18. Lamprier, S., Amghar, T., Levrat, B.: On evaluation methodologies for text segmentation algorithms. In: Proceedings of the 19th IEEE International Conference on Tools with Artificial Intelligence, vol. 2, pp. 11–18 (2007)
19. Lee, D.: Genres, registers, text types, domains, and styles: Clarifying the concepts and navigating a path through the bnc jungle. Language Learning & Technology **5**(3), 37–72 (2001)
20. Misra, H., Yvon, F., Cappe, O., Jose, J.: Text segmentation: A topic modeling perspective. Information Processing and Management **47**(4), 528–544 (2011)
21. Ponte, J.M., Croft, W.B.: Text segmentation by topic. In: Peters, C., Thanos, C. (eds.) ECDL 1997. LNCS, vol. 1324, pp. 113–125. Springer, Heidelberg (1997)
22. Reynar, J.: An automatic method of finding topic boundaries. In: Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics, pp. 331–333 (1994)
23. Riedl, M., Biemann, C.: Text segmentation with topic models. JLCL **27**(1), 47–69 (2012)
24. Scaiano, M., Inkpen, D.: Getting more from segmentation evaluation. In: Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 362–366 (2012)
25. Smolyanina, E.A.: Cohesion types in the scientific text (based on english article by m. black "metaphor"). Vestnik of Perm State University: Russian and Foreign Philology **4**(24), 140–150 (2004)
26. Stark, H.: What do paragraph markings do? Discourse Processes **11**, 275–303 (1988)
27. Trofimova, G.K.: Russian language ant the culture of speech: lectures. Flinta, Nauka (2004)
28. Tuzov, V.A.: Computer semantics of the Russian language. Saint-Petersburg University Press (2004)
29. Wan, X.: On the effectiveness of subwords for lexical cohesion based story segmentation of chinese broadcast news. Information Sciences **177**, 3718–3730 (2007)
30. Ye, N., Zhu, J., Wang, H., Ma, M., Zhang, B.: An improved model of dotplotting for text segmentation. Journal of Chinese Language and Computing **17**(1), 27–40 (2007)