TEXT SEGMENTATION TECHNIQUES: A CRITICAL REVIEW

Irina Pak and Phoey Lee Teh

Department of Computing and Information Systems, Sunway University, Bandar Sunway, Malaysia
I.Pak
e-mail: irina.p@imail.sunway.edu.my
P.L. Teh
e-mail: phoeyleet@sunway.edu.my

Department of Computing and Information Systems,

Sunway University, Bandar Sunway, Malaysia

irina.p@imail.sunway.edu.my, phoeyleet@sunway.edu.my

Abstract Text segmentation is widely used for processing text. It is a method of splitting a document into smaller parts, which is usually called segments. Each segment has its relevant meaning. Those segments categorized as word, sentence, topic, phrase or any information unit depending on the task of the text analysis. This study presents various reasons of usage of text segmentation for different analyzing approaches. We categorized the types of documents and languages used. The main contribution of this study includes a summarization of 50 research papers and an illustration of past decade (January 2007- January 2017)'s of research that applied text segmentation as their main approach for analysing text. Results revealed the popularity of using text segmentation in different languages. Besides that, the "word" seems to be the most practical and usable segment, as it is the smaller unit than the phrase, sentence or line.

1 Introduction

Text segmentation is process of extracting coherent blocks of text [1]. The segment referred as “segment boundary” [2] or passage [3]. Another two studies referred segment as subtopic [4] and region of interest [5]. There are many reasons why the splitting document can be useful for text analysis. One of the main reasons is because they are smaller and more coherent than whole documents [3]. Another reason is each segment is used as units of analysis and access [3]. Text segmentation was used to process text in emotion extraction [6], sentiment mining [7][8], opinion mining [9][10], topic identification [11][12], language detection [13] and information retrieval [14]. Sentiment analyzing within the text covers wide range of techniques, but most of them include segmentation stage in text process. For instance, Zhu et al. [15] used segmentation in his model to identify multiple polarities and aspects within one sentence. There are studies that applied tokenization in the semantic analysis to increase the probability of obtaining the useful information by processing tokens. For example, the study of Gan at al. [16] applied tokenization in their proposed method where semantics were used to improve search results for obtaining more relevant and clear content from the search. Later, Gan and Teh [17] used technique similar to segmentation where information is organized into segments called facet and values in order to improve search algorithm. Another study of Duan at al. [18] applied text segmentation and then tokenization to determine aspects and associated features. In other words, tokenization is also text segmentation because they are apparently a similar process. That is splitting text into words, symbols, phrases, or any meaningful units named as a token.

This paper reviews different methods and reasons of applying text segmentation in opinion and sentiment mining, language detection and information retrieval. The target is to overview of text segmentation techniques with brief details. The contribution of this paper includes the categorizations of recent articles and visualization of the recent trend of research in the opinion mining and related areas, such as sentiment

analysis and emotion detection. Also in pattern recognition, language processing, and information retrieval. Next section of this paper explains the scope and method used to review the past studies. Section 3 discusses the results of summarized articles. Section 4 contains a discussion on results. Lastly, section 5 concludes this paper.

2 Review Method

The review process of the articles includes publications from the past ten years. Fifty journals and conferences in total were evaluated. These articles implemented text segmentation in their main approaches. Fifty articles are summarized in Table 1 in next section.

In order to make it clear content of the Table 1, here is breakdown the column by column. The first column of the table refers to the year of the article. Next column includes the references of study. In the following column, there is a brief description of the study. There are different types of segments used in their study, it includes: 1) topic, 2) word, 3) sentence etc. Type of segment commonly selected based on their analyzing targets and specifications. The fourth column listed the type of segment used in the assessed study. The evaluation of each study has applied different sets of data or documentations. They can be categorised as: 1) corpus, 2) news, 3) articles, 4) reviews, 5) datasets etc. We listed it in the fifth column. The sixth column describes the reason for applying the text segmentation study. The last column specifies those language(s) used in those sets of documentations(s).

3 Results

Following, we present the summarized review in Table 1.

Table 1. Summarisation of the fifty articles applied text segmentation in their studies.

No	Year	Ref	Description	Segment	Data/ Document	Reason	Language
1	2007	[10]	Opinion search in web blogs (logs)	Topic	Web blog	To identify opinion in web blogs(logs)	English
2	2007	[12]	Unsupervised methods of topical text segmentation for Polish	Topic	Dataset of e-mail newsletter, artificial documents, streams, and individual documents	To detect boundaries between news items or consecutive stories	Polish
3	2007	[19]	Word segmentation of handwritten text using supervised classification techniques	Word	Dataset	To identify words within handwritten text	English
4	2007	[20]	Clustering based text segmentation	Sentence	Corpus of articles	To understand the sentences relations with consideration the similarities in a group rather than individually	English
5	2007	[6]	Comprehensive information based semantic orientation identification	Word	Comments	To identify polarity orientation in text	Chinese
6	2007	[21]	Unconventional word segmentation in Brazilian children's early text production	Word	Handwritten texts	To define word boundaries	English

7	2007	[22] Comparative Analysis of Different Text Segmentation Algorithms on Arabic News Stories	Sentence	Dataset of news	To identify boundaries within the text and measure lexical cohesion between textural units	Arabic
8	2008	[23] Automatic Story Segmentation in Chinese Broadcast News	Topic	News	To identify story boundary	Chinese
9	2009	[15] Aspect-based sentence segmentation for sentiment summarization	Sentence	Reviews	To extract aspect-based sentiment summary	Chinese
10	2009	[24] Chinese text sentiment classification based on Granule Network Zhang	Word	Corpus	To classify text based on sentiment value	Chinese
11	2009	[25] Automatic extraction of new words based on Google News corpora for supporting lexicon-based Chinese word segmentation systems	Word	Corpus of Google News	To identify word and sentence from Chinese language texts in real-word applications	Chinese
12	2010	[26] An information-extraction system for Urdu-a resource-poor language	Word	Blogs, comments, news articles	To identify social and human behavior within text	Urdu
13	2010	[27] Chinese text segmentation: A hybrid approach using transductive learning moreover, statistical association measures	Word	Corpus	To build system for cross-language information	Chinese

14	2010	[7] Sentiment classification for stock news	Word	Chinese stock news	To classify news by sentiment orientation	Chinese
15	2010	[8] Sentiment text classification of customers reviews on the web based on SVM	Word	Comments	To improve accuracy of sentiment classification	Chinese
16	2010	[28] The application of text mining technology in monitoring the network education public sentiment	Word	Web documents	To analyze sentiment in text to monitor public network	Chinese
17	2010	[29] Using text mining and sentiment analysis for online forums hotspot detection and forecast	Word	Forums	To design a text sentiment approach	Chinese
18	2011	[30] A topic modeling perspective for text segmentation.	Topic	Corpus	To design an enhance topic extraction approach	English
19	2011	[4] Iterative approach to text segmentation	Topic Subtopic	Articles	To identify topic and subtopic boundaries within content	English
20	2011	[31] Text segmentation of consumer magazinesblocks in PDF format	Text	Articles in PDF documents	To process PDF documents	English
21	2011	[32] Rule-based Malay text segmentation tool	Sentence	Articles	To design Malay sentence splitter and tokenizer	Malay English
22	2011	[33] A novel evaluation method for	Word	Corpus	To improve word segmentation	English

		morphological segmentation			with consideration true morphological ambiguity	
23	2011	[34] Usage of text segmentation and inter-passage similarities	Passage	Articles	To improve text document clustering	English
24	2012	[35] Using a boosted tree classifier for text segmentation in hand-annotated documents	Word	Dataset of handwritten documents	To convert handwritten document into digital text	English
25	2012	[1] Two-part segmentation of text documents	Word Sentence	Corpus	To process problem and solution documents	English
26	2012	[36] Enhancing lexical cohesion measure with confidence measures, semantic relations and language model interpolation for multimedia spoken content topic segmentation	Topic	Corpus of news	To split into segments that deal with a single topic	English
27	2012	[37] Text line segmentation in historical documents	Line	Historical documents	To combine the strengths of top-down and bottom-up approaches	Ancient
28	2012	[38] Topic segmentation of Chinese text based on lexical chain	Topic	News	To improve method of processing text	Chinese
29	2013	[39] Semantic-based text block segmentation	Text blocks	Web page	To retrieve image based on text around it	English

30	2013	[40] The first steps in developing machine translation of patents	Word	Dataset	To archive the translation from English to Russian	Russian English
31	2013	[41] Segmentation system based on the sentiments expressed in the text.	Tag Word	Reviews	To design system which identifies a sentiment expressed in text	English
32	2013	[42] Probabilistic Chinese word segmentation with non-local information and stochastic training	Word Character	Corpus	To reduce computational complexity of learning non-local information	Chinese
33	2013	[43] Unknown Chinese word extraction based on variety of overlapping strings	Word	Corpus	To extract words from a sentence and improve extraction of unknown words	Chinese
34	2013	[44] Text segmentation for language identification in Greek forums	Sentence Topic	Forums	To identify language	Greek English
35	2014	[13] Recognition-based segmentation of online Arabic text recognition	Word	Dataset	To recognize Arabic text within handwriting	Arabic
36	2014	[45] Word segmentation of overlapping ambiguous strings during Chinese reading	Character	Collected dataset	To detect word boundaries and recognize the word	Chinese
37	2015	[46] Chinese text sentiment orientation identification	Character	Corpus	To identify sentiment in the text	Chinese

38	2015	[47] Text segmentation based on semantic word embedding	Word	Articles	To enhance semantic word embedding approach	English
39	2015	[48] Dynamic non-parametric joint sentiment topic mixture model	Word	Dataset of forums	To filter out low/high-frequency words, single words, improper characters and stop words	Chinese
40	2015	[49] A multi-label classification based approach for sentiment classification	Word	Microblogs	To support a sentiment classification approach	Chinese
41	2015	[50] Topic segmentation of TV-streams by watershed transform and vectorization	Topic	Dataset of TV streams	To enhance and a topic extraction	French
42	2015	[51] A supervised fine-grained sentiment analysis system for online reviews	Word	Corpus of hotel reviews	To pre-process data for sentiment analysis	Chinese
43	2016	[52] Vietnamese word segmentation	Word	Dictionaries	To check how dictionary size affects word segmentation	Vietnamese
44	2016	[53] Text segmentation of digitized clinical texts	Line	Corpus	To identify column separator	French
45	2016	[54] Phrase-level segmentation and labeling	Phrase	Dataset	To balance between the word and sentence levels	English

46	2016	[55] Akkadian word segmentation	Word	Corpus	To improve the language processing in cuneiform	Ancient Akkadian
47	2016	[56] Document segmentation and classification into musical scores and text	Word	Dataset	To detect bounding boxes containing the musical score and text	English
48	2016	[57] Candidate document retrieval for cross-lingual plagiarism detection	Topic	Corpus	To convert the suspicious document to a set of related passages	English German Spanish
49	2016	[58] Effects of text segmentation on silent reading of Chinese regulated poems	Character	Chinese regulated poems	To monitor eye movements of native participants in reading Chinese regulated poems	Chinese
50	2017	[59] A new watershed model-based system for character segmentation in degraded text lines	Character	Dataset of historical document	To understand the content in historical documents	Indian English South Indian

Table 1 presents the summarization for the review of past years' studies where different types of segmentation are specified. As well as the reason for applying text segmentation in their methods. Doubtless, the main reason of the researchers using any text segmentation in text processing is splitting documents into segments. After the document is split, segments proceed to the different phase depending on the text analysis approach.

In order to summarize and evaluate Table 1 in more details, pie charts and graphs are presented below. Each of the charts help to look at presented papers from different perspectives. Figure 1 concentrates on types of segment used in the studies. It is essential to evaluate segments' categorization as it is one of the main aim of this paper to consider segmentation features. Figure 2 presents a pie chart which shows variety of different languages used in evaluated papers. Lastly, Figure 3 presents a graph which indicates types of segments used in different documents' types. That can help to get a picture of which type of documents are highly used and see the another trend regarding types of segment.

Figure 1 visualizes ten different type of segments used in text segmentation from the fifty articles. Word segment is highly used compare to other types of segment. It occupied the biggest area of the chart, equivalent to a total of 26 from 50 reviewed articles. There can be several reasons to use this technique. For instance, Homburg and Chiarcos [55] described it as the most elementary and essential task in natural language processing of written language. Character segment can be categorized as word segment as well because it is mostly applied in Chinese text processing, as character segment represents single unit same as word segment in following studies[45][46][58]. Another study of [56] proposes a method for segmentation of musical documents using word segment. It detects the segments and assigns each segment to particular musical score or text.

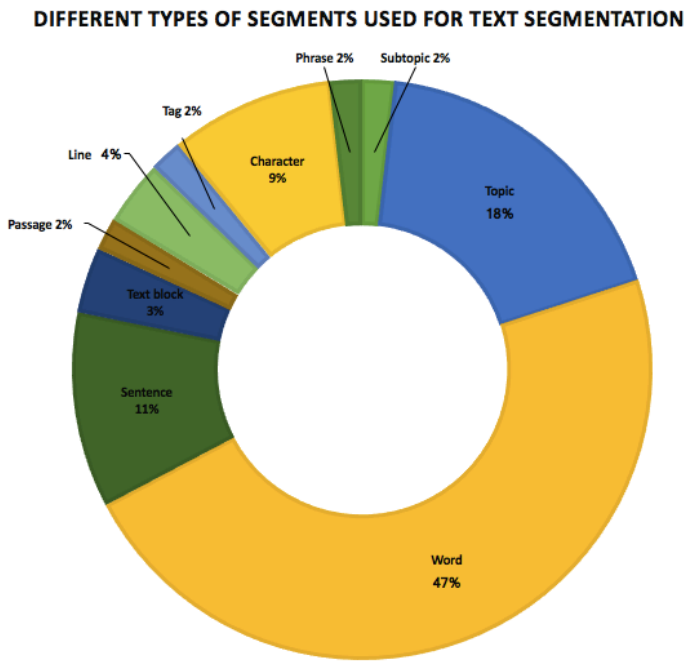

Fig. 1. Percentage of different used for text segmentation.

The second biggest area of the chart covers the use of topic segment in text processing. Topic segmentation plays an important role in data processing. For example, topic segmentation is successfully applied in tackling the problem of information overload that occur when the whole document is presented at once. Misra et al. [30] stated the reason behind splitting document could be reasonable to present only the relevant part(s) of a document. The reason they stated is that presenting the whole document without segmentation may result in previously discussed information overload. Paliwal and Pudi [34] also addressed the same problem, which led them to propose a clustering approach based on topic segmentation. Topic segmentation is popular in opinion mining area. For instance, studies of [10][30][38] used the topic as a segment. Another example of topic segmentation applied in opinion mining is a study of Claveau and Lefevre [50]. They proposed a new technique to compute similarities between candidate segments. Two corpora of TV broadcast in

French used in evaluating of proposed technique. Furthermore, a study by Song et. al. [4] has used the topic as segment too; they proposed a novel method that includes hierarchical organization and language modeling to split the text into parts. The result of that study showed that proposed method is effective in identifying the topics in evaluated dataset of articles. Apart from opinion mining, topic segmentation is used in cross-language plagiarism detection. For instance, Ehsan and Shakery [57] applied to find and examine the candidate retrieval, where proposed approach converts the document to a set of related passages. After that, it uses a proximity-based model to retrieve documents with the best matching passages. The third most applied type of segment is a sentence. The common issue in sentence segmentation among assessed articles is identifying the boundaries between sentences [22][20].

From Table 1, it is obvious that text segmentation was applied to process text in a variety of languages including English. The pie chart on Figure 2 illustrates the numbers of percentage for each language used including English from Table 1, in order to see trend of languages used. Leading language among evaluated articles is English with the result of 38%.

Besides English, the Chinese language is the most widely evaluated. This result shows the highest use of word segmentation technique in the process. In short, studies of [6][24][7][8][27][28][29][43][46][48][49][58] used word or character segmentation, and they are all analyzing Chinese text. One of them is study of Xia et al. [24] presents a new approach for Chinese sentiment classification based on granule network. They applied word segmentation to split sentences and later select sentimental candidate words.

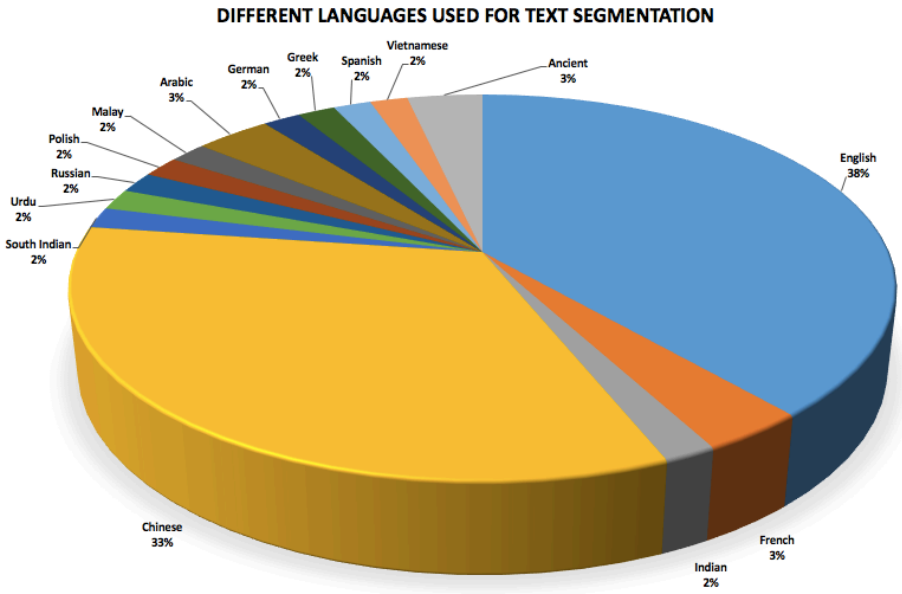

Fig. 2. Percentage of different languages used for text segmentation.

Another study of Lan at al.[46] came up with another way of analyzing Chinese text in the same area of sentiment classification. Instead of using word text segmentation, they extract sentiment value based on each character, claiming that each character can contain rich sentiment information. Another study by Hog at al. [25] applied Chinese word segmentation in information retrieval process. They built an automatic statistics-based scheme for extracting news word based on the corpora. One of the main useful features of the proposed scheme is automatic and enhanced word identification. Beside the Chinese language, there are studies which applied word segmentation for Urdu [26], Russian [40], Arabic [13], Vietnamese [52], Akkadian [55], Indian [59] and South Indian [59]. However, other studies [22][44] and [32] applied sentence segmentation to analyze Arabic, Greek, and Malay languages accordingly. The topic segmentation technique is used for processing Polish [10], Chinese [38], French [50], German [57] and Spanish [57] languages. Besides that, “line” segmentation is also applied for processing Ancient [37] and French [53] languages. As a summary, it can be seen that there is

a trend to apply text segmentation in the analyzing text in different languages.

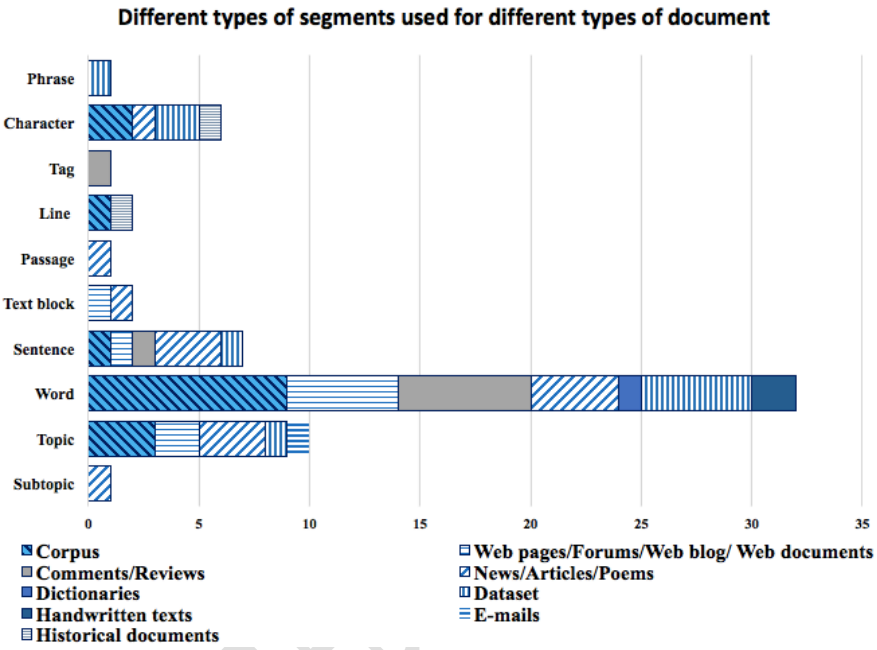

Fig. 3. Different types of segments used for different types of data.

We further identify if the different types of the document and datasets have any relationship on the selection of different type of segmentation technique. Figure 3 presents the number of percentage of each type of segments used for different types of document derived from Table 1. In this study, we categorize web pages, web blog, and web documents as the same group of documentations. After all, it is the most used type of document among all. The second widely used type of document is comments and reviews. Segments can be classified into the tag, word, sentence, topic, phrase and any information unit.

Figure 3 concludes that word is the most used type of segment used for corpus. Furthermore, the word segment is also highly used for comments and reviews. Web documents are utilized under word, topic and sentence segments that can indicate that web documents including pages, forums and blogs take a big part in text analyzing.

4 Discussion

As a result, we noticed the trend of applying text and sentence segmentation in processing and analyzing different languages such as Chinese, Vietnamese, Urdu, Arabic, and Ancient languages. Besides applying text segmentation for different languages, text segmentation successfully applied in opinion mining for news, blog and stock market. Finally, after comparison between evaluated types of segments, word segment is the most used compare to another types of the segment. It can be due to the smallest size of the segment which allows more detailed analysis.

5 Conclusion

This paper presents a critical review of the text segmentation methods and reasons in text processing and analyzing languages, sentiment, opinions and fifty published articles for the past decade were categorized and summarized. Those articles give contributions to text processing in information retrieval, emotion extraction, sentiment and opinion mining and language detection.

Results of this study show that word as the segment is the most used compare to other types of the segment. It means that processing smaller segments can be more useful and meaningful for more detailed and deeper analyzing of the text. Different types of document are used as a dataset for the experiment. The most popular are web pages, web blog and web document following by comments and reviews. That indicates

that information from the online users and consumers plays an important role in expressing people's emotions, opinions, and feelings.

Considering the findings of this paper, the future study can include implementation of text analysis approach using text segmentation with word segment.

Acknowledgments. We would like to thank First EAI International Conference on Computer Science and Engineering for the opportunity to present our paper and further extend it. This research paper was partially supported by Sunway University Internal Research Grant No. INT-FST-IS-0114-07 and Sunway-Lancaster Grant SGSSL-FST-DCIS-0115-11.

References

1. P. D, Visweswariah K, Wiratunga N, Sani S (2012) Two-part segmentation of text documents. In: Proc. 21st ACM Int. Conf. Inf. Knowl. Manag. - CIKM '12. ACM New York, NY, Maui, p 793
2. Scaiano M, Inkpen D, Laganière R, Reinhartz A (2010) Automatic text segmentation for movie subtitles. In: Lect. Notes Comput. Sci. Springer, pp 295–298
3. Oh H, Myaeng SH, Jang M-G (2007) Semantic passage segmentation based on sentence topics for question answering. Inf Sci (Ny) 177:3696–3717.
4. Song F, Darling WM, Duric A, Kroon FW (2011) An Iterative Approach to Text Segmentation. In: 33rd Eur. Conf. IR Res. ECIR 2011. Springer Berlin Heidelberg, Dublin, pp 629–640
5. Oyedotun OK, Khashman A (2016) Document segmentation using textural features summarization and feedforward neural network. Appl Intell 45:1–15.
6. Wu Y, Zhang Y, Luo SM, Wang XJ (2007) Comprehensive information based semantic orientation identification. In: IEEE NLP-KE 2007 - Proc. Int. Conf. Nat. Lang. Process. Knowl. Eng. IEEE, Beijing, pp 274–279
7. Gao Y, Zhou L, Zhang Y, et al (2010) Sentiment classification for stock news. In: ICPCA10 - 5th Int. Conf. Pervasive Comput. Appl. IEEE, Maribor, pp 99–104
8. Xia H, Tao M, Wang Y (2010) Sentiment text classification of customers reviews on the Web based on SVM. Proc - 2010 6th Int Conf Nat Comput ICNC 2010 7:3633–3637.
9. Liu C, Wang Y, Zheng F (2006) Automatic text summarization for dialogue style. In: Proc. IEEE ICIA 2006 - 2006 IEEE Int. Conf. Inf. Acquis. IEEE, Weihai, pp 274–278
10. Osman DJ, Yearwood JL (2007) Opinion search in web logs. Conf Res Pract Inf Technol Ser 63:133–139.
11. Brants T, Chen F, Tsochantaridis I (2002) Topic-Based Document Segmentation with Probabilistic Latent Semantic Analysis. In: CIKM'02. ACM, Virginia, pp 211–218
12. Flejter D, Wieloch K, Abramowicz W (2007) Unsupervised Methods of Topical Text Segmentation for Polish. In: SIGIR'13. ACM, Dublin, pp 51–58
13. Potrus MY, Ngah UK, Ahmed BS (2014) An evolutionary harmony search algorithm with dominant point detection for recognition-based segmentation of online Arabic text recognition. Ain Shams Eng J 5:1129–1139.
14. Huang X, Peng F, Schuurmans D, et al (2003) Applying Machine Learning to Text Segmentation. Inf Retr J 6:333–362.

15. Zhu J, Zhu M, Wang H, Tsou BK (2009) Aspect-based sentence segmentation for sentiment summarization. In: Proceeding 1st Int. CIKM Work. Top. Anal. mass Opin. - TSA '09. ACM New York, NY, USA ©2009, Hong Kong, pp 65–72
16. Gan KH, Phang KK, Tang EK (2007) A semantic learning approach for mapping unstructured query to web resources. In: Proc. - 2006 IEEE/WIC/ACM Int. Conf. Web Intell. (WI 2006 Main Conf. Proceedings), WI'06. IEEE, Hong Kong, pp 494–497
17. Hoon GK, Wei TC (2016) Flexible Facets Generation for Faceted Search. In: First EAI Int. Conf. Comput. Sci. Eng. EAI, Penang, Malaysia, pp 1–3
18. Duan D, Qian W, Pan S, et al (2012) VISA: a Visual Sentiment Analysis System. In: Proc. 5th Int. Symp. Vis. Inf. Commun. Interact. - VINCI '12. ACM, Hangzhou, pp 22–28
19. Sun Y, Butler TS, Shafarenko A, et al (2007) Word segmentation of handwritten text using supervised classification techniques. Appl Soft Comput 7:71–88.
20. Lamprier S, Amghar T, Levrat B, Saubion F (2007) ClassStruggle: a Clustering Based Text Segmentation. In: Proc. SAC '07. ACM, Seoul, pp 600–604
21. Correa J, Dockrell JE (2007) Unconventional word segmentation in Brazilian children's early text production. Read Writ 20:815–831.
22. El-Shayeb MA, El-Beltagy SR, Rafea A (2007) Comparative Analysis of Different Text Segmentation Algorithms on Arabic News Stories. In: IEEE Int. Conf. Inf. Reuse Integr. Las Vegas, pp 441–446
23. Xie L, Zeng J, Feng W (2008) Multi-Scale TextTiling for Automatic Story Segmentation in Chinese Broadcast News. In: 4th Asia Infomation Retr. Symp. Springer Berlin Heidelberg, Harbin, pp 345–355
24. Xia Z, Suzhen W, Mingzhu X, Yixin Y (2009) Chinese text sentiment classification based on granule network. In: 2009 IEEE Int. Conf. Granul. Comput. GRC 2009. IEEE, Nanchang, pp 775–778
25. Hong CM, Chen CM, Chiu CY (2009) Automatic extraction of new words based on Google News corpora for supporting lexicon-based Chinese word segmentation systems. Expert Syst Appl 36:3641–3651.
26. Mukund S, Srihari R, Peterson E (2010) An Information-Extraction System for Urdu-A Resource-Poor Language. ACM Trans Asian Lang Inf Process 9:1–43.
27. Tsai RT-H (2010) Chinese text segmentation: A hybrid approach using transductive learning and statistical association measures. Expert Syst Appl 37:3553–3560.
28. Liu X, Zuo M, Chen L (2010) The application of text mining technology in monitoring the network education public sentiment. In: 2010 Int. Conf. Comput. Intell. Softw. Eng. IEEE, Wuhan, pp 1–4
29. Li N, Wu DD (2010) Using text mining and sentiment analysis for online forums hotspot detection and forecast. Decis Support Syst 48:354–368.
30. Misra H, Yvon F, Cappé O, Jose J (2011) Text segmentation : A topic modeling perspective. Inf Process Manag 47:528–544.
31. Fan J (2011) Text segmentation of consumer magazines in PDF format. Int Conf Doc Anal Recognition, ICDAR 794–798.
32. Ranaivo-Malançon B (2011) Building a rule-based Malay text segmentation tool. In: 2011 Int. Conf. Asian Lang. Process. IALP 2011. IEEE, Penang, pp 276–279
33. Nouri J, Yangarber R (2011) A Novel Evaluation Method for Morphological Segmentation. In: Proc. Tenth Int. Conf. Lang. Resour. Eval. (LREC 2016). European Language Resources Association (ELRA), Portoroz, pp 3102–3109
34. Paliwal S, Pudi V (2012) Investigating Usage of Text Segmentation and Inter-passage Similarities. In: Mach. Learn. Data Min. Pattern Recognit. Springer Berlin Heidelberg, Berlin, pp 555–565
35. Peng X, Setlur S, Govindaraju V, Ramachandrula S (2012) Using a boosted tree classifier for text segmentation in hand-annotated documents. Pattern Recognit Lett 33:943–950.

36. Guinaudeau C, Gravier G, S??billot P (2012) Enhancing lexical cohesion measure with confidence measures, semantic relations and language model interpolation for multimedia spoken content topic segmentation. Comput Speech Lang 26:90–104.
37. Clausner C, Antonacopoulos A, Pletschacher S (2012) A robust hybrid approach for text line segmentation. In: 21st Int. Conf. pattern Recognit. IEEE, Tsukuba, pp 335–338
38. Ye FY, Chen Y, Luo X, et al (2012) Research on topic segmentation of Chinese text based on lexical chain. In: 12th Int. Conf. Comput. Inf. Technol. CIT 2012. IEEE, Chengdu, pp 1131–1136
39. Myint N, Aung M, Maung SS (2013) Semantic Based Text Block Segmentation Using WordNet. Int J Comput Commun Eng 2:601–604.
40. Kravets LG (2013) The first steps in developing machine translation of patents. World Pat Inf 35:183–186.
41. Chiru C, Teka A (2013) Sentiment-Based Text Segmentation. In: 2nd Int. Conf. Syst. Comput. Sci. IEEE, Villeneuve d'Ascq, France, pp 234–239
42. Sun X, Zhang Y, Matsuzaki T, et al (2013) Probabilistic Chinese word segmentation with non-local information and stochastic training. Inf Process Manag 49:626–636.
43. Ye Y, Wu Q, Li Y, et al (2013) Unknown chinese word extraction based on variety of overlapping strings. Inf Process Manag 49:497–512.
44. Fragkou P (2013) Text Segmentation for Language Identification in Greek Forums. In: Proc. Adapt. Lang. Resour. Tools Closely Relat. Lang. Lang. Var. Elsevier B.V., Hissar, pp 23–29
45. Ma G, Li X, Rayner K (2014) Word segmentation of overlapping ambiguous strings during Chinese reading. J Exp Psychol Hum Percept Perform 40:1046–1059.
46. Lan Q, Li W, Liu W (2015) Chinese Text Sentiment Orientation Identification Based on Chinese-Characters. In: Fuzzy Syst. Knowl. Discov. (FSKD), 2015 12th Int. Conf. IEEE, Zhangjiajie, pp 663–668
47. Alemi AA, Ginsparg P (2015) Text Segmentation based on Semantic Word Embeddings. In: KDD2015. ACM, Sydney, Australia, pp 1–10
48. Fu X, Yang K, Huang JZ, Cui L (2015) Dynamic non-parametric joint sentiment topic mixture model. Knowledge-Based Syst 82:102–114.
49. Liu SM, Chen J-H (2015) A multi-label classification based approach for sentiment classification. Expert Syst Appl 42:1083–1093.
50. Claveau V, Lefevre S (2015) Topic segmentation of TV-streams by watershed transform and vectorization. Comput Speech Lang 29:63–80.
51. Shi H, Zhan W, Li X (2015) A Supervised Fine-Grained Sentiment Analysis System for Online Reviews. Intell Autom Soft Comput 21:589–605.
52. Liu W, Wang L (2016) How does Dictionary Size Influence Performance of Vietnamese Word Segmentation ? In: Proc. Tenth Int. Conf. Lang. Resour. Eval. (LREC 2016). European Language Resources Association (ELRA), Portorož, Slovenia, pp 1079–1083
53. Grouin C (2016) Text segmentation of digitized clinical texts. In: Proc. Tenth Int. Conf. Lang. Resour. Eval. (LREC 2016). European Language Resources Association (ELRA), Portorož, Slovenia, pp 3592–3599
54. Logacheva V, Specia L (2016) Phrase-Level Segmentation and Labelling of Machine Translation Errors. In: Tenth Int. Conf. Lang. Resour. Eval. (LREC 2016). European Language Resources Association (ELRA), Portorož, Slovenia, pp 2240–2245
55. Homburg T, Chiarcos C (2016) Akkadian Word Segmentation. In: Proc. Tenth Int. Conf. Lang. Resour. Eval. (LREC 2016). European Language Resources Association (ELRA), Portorož, Slovenia, pp 4067–4074
56. Pedersoli F, Tzanetakis G (2016) Document segmentation and classification into musical scores and text. Int J Doc Anal Recognit 19:289–304.
57. Ehsan N, Shakery A (2016) Candidate document retrieval for cross-lingual plagiarism detection using two-level proximity information. Inf Process Manag 52:1004–1017.

58. Qingrong C, Wentao G, Scheepers C, et al (2017) Effects of text segmentation on silent reading of Chinese regulated poems : Evidence from Eye Movements. 44:265–286.
59. Kavitha AS, Shivakumara P, Kumar GH, Lu T (2017) A new watershed model based system for character segmentation in degraded text lines. AEU - Int J Electron Commun 71:45–52.