Text Segmentation for Analysing Different Languages

Irina Pak and Phoey Lee Teh

Department of Computing and Information Systems,
Sunway University, Bandar Sunway, Malaysia
irina.p@imail.sunway.edu.my, phoeyleet@sunway.edu.my

Abstract. Over the past several years, researchers have applied different methods of text segmentation. Text segmentation is defined as a method of splitting a document into smaller segments, assuming with its own relevant meaning. Those segments can be classified into the tag, word, sentence, topic, phrase and any information unit. Firstly, this study reviews the different types of text segmentation methods used in different types of documentation, and later discusses the various reasons for utilizing it in opinion mining. The main contribution of this study includes a summarisation of research papers from the past 10 years that applied text segmentation as their main approach in text analysing. Results show that word segmentation was successfully and widely used for processing different languages.

Keywords: Text Segmentation, Text Analysis, Text Processing, Languages, Online Reviews, Opinion Mining.

1 Introduction

Segmentation is splitting a document into segments. The segment is also referred as "segment boundary" [1] or passage [2]. While another studies referred segment as subtopic [3] and region of interest [4]. There are many reasons why the splitting document can be useful for text analysis. One of the main reasons is handling those segments is easier since they are smaller and more coherent than whole documents [2]. Another reason is segments can be used as units of analysis and access [2]. Text segmentation was used to process text in opinion mining [5] [6], information retrieval [7], emotion extraction [8], sentiment mining [9] [10] and language detection [11].

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

Section 2 of this paper explains the method used to review the past studies. Section 3 discusses the results of summarised articles. Section 4 concludes this paper.

2 Review Method

This paper is limited to thirty articles which are summarised in Table 1. It includes a summarization of past 10 years of articles from journals and conferences that involved text segmentation as their approaches. The first column contains the references [12-35] of the articles. The second column contains the year of the evaluated study. Following column briefly, describes the evaluated study. The fourth column states the type of segment used in the evaluated study. The segmentation is used for different types of document, column five contains information regarding the type of documents. The reasons for applying segmentation in particular studies are stated in the following column. The last column shows the type of language in tested documents.

3 Results

Table 1. Summarisation of articles.

References	Year	Description	Type of segment	Type of documents	Reason	Language
[12]	2006	Improving text segmentation using latent semantic analysis	Sentence	Corpus	To improve the accuracy of segmentation	Belgian French
[5]	2006	Automatic text summarization for dialogue style	Text block	Dialogue samples Corpus	To identify dialogue content	Chinese
[13]	2006	HTML text segmentation for Web page summarization	Sentence	Web pages	To identify Web page content	Japanese
[6]	2007	Opinion search in web blogs (logs)	Topic	Web blog	To identify opinion in web blogs(logs)	English
[8]	2007	Comprehensive information based semantic orientation identification	Word	Comments	To identify polarity orientation in text	Chinese
[14]	2008	Automatic Story Segmentation in Chinese Broadcast News	Topic	News	To identify story boundary	Chinese
[15]	2009	Aspect-based sentence segmentation for sentiment summarization	Sentence	Reviews	To extract aspect-based sentiment summary	Chinese
[16]	2009	Chinese text sentiment classification	Word	Corpus	To classify text based on sentiment value	Chinese
[17]	2010	An information-extraction system for Urdu-a resource-poor language	Word	Blogs, comments, news articles	To identify social and human behaviour within text	Urdu
[9]	2010	Sentiment classification for stock news	Word	Chinese stock news	To classify news by sentiment orientation	Chinese
[10]	2010	Sentiment text classification of customers reviews on the web based on SVM	Word	Comments	To improve accuracy of sentiment classification	Chinese
[18]	2010	The application of text mining technology in monitoring the network education public sentiment	Word	Web documents	To analyse sentiment in text to monitor public network	Chinese
[19]	2010	Using text mining and sentiment analysis for online forums hotspot detection and	Word	Forums	To design a text sentiment approach	Chinese

		forecast				
[20]	2011	A topic modelling perspective for text segmentation.	Topic	Corpus	To design an enhanced topic extraction approach	English
[21]	2011	Text segmentation of consumer magazines in PDF format	Text blocks	Articles in PDF documents	To process PDF documents	English
[22]	2011	Rule-based Malay text segmentation tool	Sentence	Articles	To design Malay sentence splitter and tokenizer	Malay English
[23]	2012	Usage of text segmentation and inter-passage similarities	Passages	Articles	To improve text document clustering	English
[24]	2012	Two-part segmentation of text documents	Word Sentence	Corpus	To process problem and solution documents	English
[25]	2012	Text line segmentation in historical documents	Line	Historical documents	To combine the strengths of top-down and bottom-up approaches	Ancient
[26]	2012	Topic segmentation of Chinese text based on lexical chain	Topic	News	To improve method of processing text	Chinese
[27]	2013	Semantic-based text block segmentation	Text blocks	Web page	To retrieve image based on text around it	English
[28]	2013	Segmentation system based on the sentiments expressed in the text.	Tag Word	Reviews	To design system which identifies a sentiment expressed in text/	English
[11]	2014	Recognition-based segmentation of online Arabic text recognition	Word	Dataset	To recognise Arabic text within handwriting	Arabic
[29]	2014	Text segmentation for language identification in Greek forums	Sentence Topic	Forums	To identify language	Greek English
[30]	2015	Chinese text sentiment orientationr identification	Characte	Corpus	To identify sentiment in the text	Chinese
[31]	2015	Text segmentation based on Semantic word embedding	Word	Articles	To enhance semantic word embedding approach	English
[32]	2015	A multi-label classification based approach for sentiment classification	Word	Microblogs	To support a sentiment classification approach	Chinese
[33]	2016	Vietnamese word segmentation	Word	Dictionaries	To check how dictionary sizeese affects word segmentation	Vietnam

[34]	2016	Phrase-level segmentation and labelling	Phrase	Dataset	To balance English between the word and sentence levels
[35]	2016	Akkadian word segmentation	Word	Corpora	To improve the Ancient language processingAkkadian in cuneiform

Table 1 presents the summarisation for the review of past years' studies. Different types of segmentation are discussed. For instance, topic segmentation has been successfully applied in tackling the problem of information overload that occur when the whole document is presented at once. Misra et al. [20] stated the reason behind splitting document can be reasonable to present only the relevant part(s) of a document, because presenting the whole document without segmentation may result in information overload. Paliwal and Pudi [23]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 [20] [6] [26] used the topic as a segment.

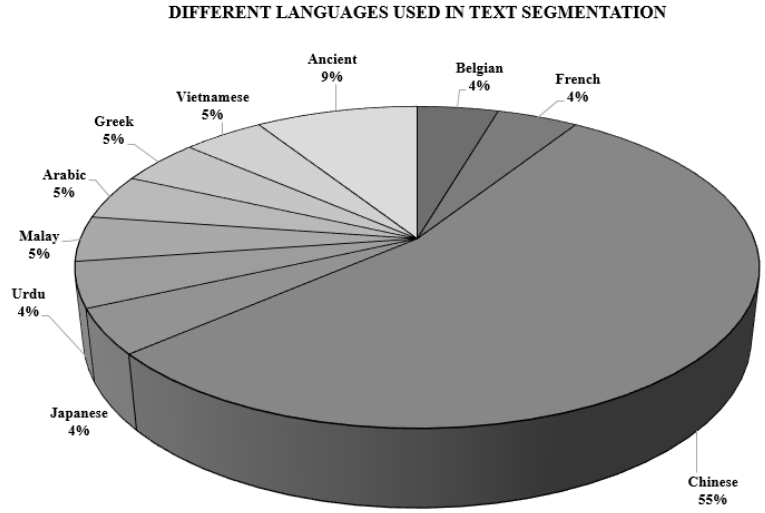

Fig. 1. Percentage of different languages used for text segmentation.

Another type of segmentation is word segmentation, Homburg and Chiarcos [35] describe word segmentation as the most elementary task in natural language processing of written language. This type of segmentation was applied in language detection. Figure 1 illustrates the numbers of percentage for each language beside English used in evaluated studies. For instance, studies of [5] [8] [9] [10] [16] [18] [19] [30] [32] used word (character) segmentation for analysing Chinese text. Beside the Chinese language, there are studies which applied word segmentation for Urdu[17], Arabic [11], Vietnamese [33], and Akkadian [35]. However, other studies [13][29][22] applied sentence segmentation to analyse Japanese, Greek and Malay languages accordingly. As it is seen, there is a trend to apply text segmentation in the analysing text in different languages.

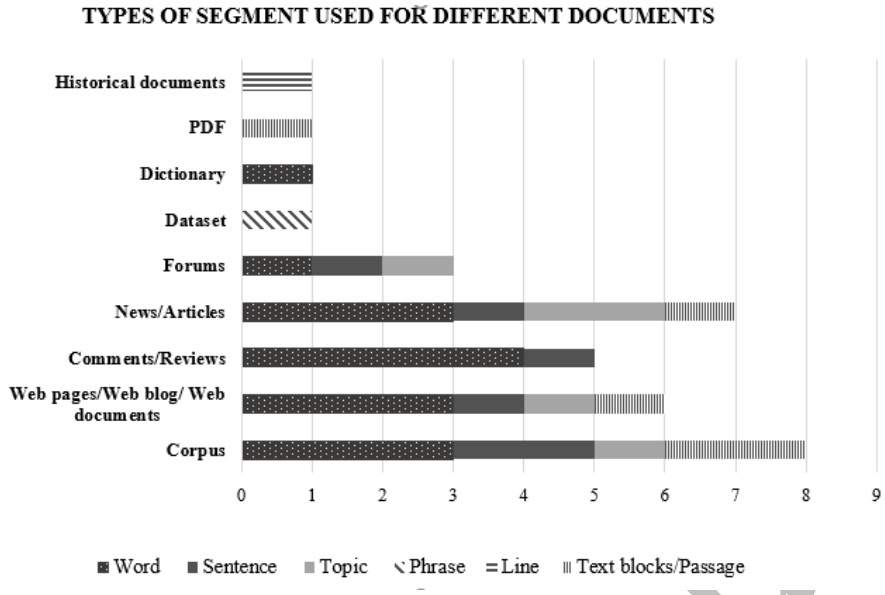

Fig. 2. Chart of different types of segments used for different types of document.

Different types of the document and datasets were used in research experiments in order to check text segmentation accuracy. Figure 2 presents a number of each type of segments used for different types of document derived from Table 1. In this study, web pages, web blog and web documents are categorised as one group. By comparison, it is the most used type of document referring to the Figure 2. After web pages, web blog and web documents, 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 2 concludes that word is the most used type of segment.

As a result, we noticed the trend of applying text and sentence segmentation in processing and analysing different languages such as Chines, 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, word segment is the most used compare to another types of the segment. The reason can be an as smaller segment to process as more detailed analysis can be done.

4 Conclusion

This paper presents an overview of the text segmentation methods and reasons in text processing and analysing. Thirty published articles for past 10 years were categorised and summarised. Those articles give contributions to text processing in information retrieval, emotion extraction, sentiment mining and language detection.

Results 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 deeper and more detailed analysing 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.

References

[1] Scaiano M, Inkpen D, Laganière R, Reinhartz A (2010) Automatic text segmentation for movie subtitles. In: Lect. Notes Comput. Sci. Springer, pp 295–298

[2] Oh H, Myaeng SH, Jang M-G (2007) Semantic passage segmentation based on sentence topics for question answering. Inf Sci (Ny) 177:3696–3717.

[3] Song F, Darling WM, Duric A, Kroon FW (2011) An Iterative Approach to Text Segmentation. In: 33rd Eur. Conf. IR Res. 2011. Springer Berlin Heidelberg, Dublin, pp 629–640

[4] Oyedotun OK, Khashman A (2016) Document segmentation using textural features summarization and feedforward neural network. Appl Intell 45:1–15.

[5] 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

[6] Osman DJ, Yearwood JL (2007) Opinion search in web logs. Conf Res Pract Inf Technol Ser 63:133–139.

[7] Huang X, Peng F, Schuurmans D, et al (2003) Applying Machine Learning to Text Segmentation. Inf Retr J 6:333–362.

[8] 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

[9] 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

[10] 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.

[11] 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.

[12] Bestgen Y (2006) Improving Text Segmentation Using Latent Semantic Analysis : A Reanalysis of Choi ,. Assoc Comput Linguist 32:5–12.

[13] Sunayama W, Iyama A, Yachida M (2006) HTML text segmentation for Web page summarization by a key sentence extraction method. Syst Comput Japan 37:26–36.

[14] 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

[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, Hong Kong, pp 65–72

[16] 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

[17] 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.

[18] 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

[19] Li N, Wu DD (2010) Using text mining and sentiment analysis for online forums hotspot detection and forecast. Decis Support Syst 48:354–368.

[20] Misra H, Yvon F, Cappé O, Jose J (2011) Text segmentation : A topic modeling perspective. Inf Process Manag 47:528–544.

[21] Fan J (2011) Text segmentation of consumer magazines in PDF format. Int Conf Doc Anal Recognition, ICDAR 794–798.

[22] 22. 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

[23] 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

[24] 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, USA ©2012, Maui, p 793

[25] 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
[26] 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.IEEE, Chengdu, pp 1131–1136
[27] Myint N, Aung M, Maung SS (2013) Semantic Based Text Block Segmentation Using WordNet. Int J Comput Commun Eng 2:601–604.
[28] Chiru C, Teka A (2013) Sentiment-Based Text Segmentation. In: 2nd Int. Conf. Syst. Comput. Sci. IEEE, Villeneuve d'Ascq, France, pp 234–239
[29] 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
[30] Lan Q, Li W, Liu W (2015) Chinese Text Sentiment Orientation Identification Based on Chinese-Characters. In: Fuzzy Syst. Knowl. Discov., 2015 12th Int. Conf. IEEE, Zhangjiajie, pp 663–668
[31] Alemi AA, Ginsparg P (2015) Text Segmentation based on Semantic Word Embeddings. In: KDD2015. ACM, Sydney, Australia, pp 1–10
[32] Liu SM, Chen J-H (2015) A multi-label classification based approach for sentiment classification. Expert Syst Appl 42:1083–1093.
[33] Liu W, Wang L (2016) How does Dictionary Size Influence Performance of Vietnamese Word Segmentation ? In: Proc. Tenth Int. Conf. Lang. Resour. Eval. European Language Resources Association, Portorož, Slovenia, pp 1079–1083
[34] Logacheva V, Specia L (2016) Phrase-Level Segmentation and Labelling of Machine Translation Errors. In: Tenth Int. Conf. Lang. Resour. Eval. European Language Resources Association, Portorož, Slovenia, pp 2240–2245
Homburg T, Chiarcos C (2016) Akkadian Word Segmentation. In: Proc. Tenth Int. Conf. Lang. Resour. Eval. European Language Resources Association, Portorož,Slovenia, pp 4067–4074