JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Exploring and Evaluating Text Segmentation Techniques: A Comparative Analysis

Jagin M. Patel
M. K. Institute of Computer studies, Bharuch, Gujarat, India.

Abstract:

Text segmentation plays a crucial role in natural language processing tasks such as information retrieval, text summarization, and sentiment analysis. Various text segmentation techniques have been developed to extract meaningful units from a given text. This research paper aims to provide a comprehensive comparative analysis of different text segmentation techniques, evaluating their strengths, weaknesses, and suitability for specific applications.

Keywords: application domain of Image segmentation, implicit segmentation, explicit segmentation.

1. Introduction

In the ever-expanding area of natural language processing (NLP), effective text segmentation stands as a fundamental task, acting as the gateway to countless applications such as information retrieval, sentiment analysis, and text summarization. Text segmentation involves the process of breaking down a given text into meaningful and coherent segments, be it sentences, phrases, or paragraphs, facilitating a deeper understanding of the underlying content. This research embarks on a journey to conduct a comprehensive study of various text segmentation techniques, aiming to shed light on their diverse methodologies, strengths, and weaknesses, ultimately guiding practitioners and researchers towards more informed choices in adopting these techniques.

Text segmentation has been utilised to parse text in various applications such as opinion mining [1, 2], emotion extraction [3,4], sentiment mining [4,5], language detection [6], topic identification [7, 8] etc.

The importance of text segmentation cannot be overstated, given its pivotal role in enhancing the efficiency and accuracy of NLP systems. The choice of a suitable text segmentation technique significantly influences the performance of subsequent tasks, making it imperative to evaluate and understand the available methodologies. This research seeks to contribute to this understanding by systematically reviewing of text segmentation techniques.

As the NLP landscape continues to evolve, the demand for robust and versatile text segmentation techniques becomes increasingly pronounced. Rule-based methods, while providing explicit guidelines for segmentation, may lack adaptability in handling diverse linguistic patterns. Statistical approaches, on the other hand, leverage probabilistic models but may struggle with language structures. Machine learning and deep learning models, with their ability to learn complex patterns, offer promise but come with their own set of challenges.

2. Text Segmentation

Depending on the context, segmentation can have a variety of meanings. Segmentation implies separating the text area from the background [9] lines from a paragraph, words from lines, characters from the word, and so on. During the text segmentation process, a document is divided into logical sub-components such as text and images, lines in a paragraph, words in a line, and characters in words.

Character segmentation is complicated because of the broad diversity of fonts, continuously developing text styles, and image qualities such as poor-quality printing and poor binary images.

Characters that have been touched, overlapped, separated, or broken are important causes of segmentation faults. Furthermore, when a document contains many languages, character segmentation becomes more complex because to variances in character sizes and touching types of each language.

Segmentation is crucial in script recognition. Correct segmentation is the foundation for precise script recognition [10]. The segmentation approach used determines the technique used in later processing phases such as feature extraction and character recognition. Distinct segmentation procedures may result in distinct personalities.

Text segmentation Levels

Text segmentation may be at line, word and character level. The simplest and most basic phase in text-based image segmentation is line segmentation. It involves scanning the image horizontally, pixel by pixel, from left to right and top to bottom [11,12].

The next step of segmentation is word segmentation. It entails scanning the image vertically, pixel by pixel, from left to right and top to bottom [11,13].

Character-level segmentation involves the process of identifying and isolating individual characters within a text image or document. This segmentation method is essential for tasks such as Optical Character Recognition (OCR), where the accurate identification of each character is crucial.

3. Various Text segmentation techniques

Several methods have been developed, each with its strengths and applications. Here are various text segmentation techniques:

(i) Stochastic strategy

The stochastic strategy for text-based image segmentation is used by researchers in [12,13,14]. The stochastic technique uses a probabilistic algorithm to create nonlinear routes between overlapping text lines. Hidden Markov modelling (HMM) is used to extract these lines.

(ii) Supervised learning

Supervised learning in text segmentation involves training a model using labeled datasets where the input data (images or text) is paired with corresponding annotated ground truth, indicating the correct segmentation boundaries. The goal is for the model to learn patterns and features that distinguish different text regions within an image or document. Supervised learning in text segmentation is effective when sufficient labeled training data is available.

Various approaches to supervised learning in text segmentation have been researched. Fisher et al. [15] used a standard machine learning approach and a variety of finite-state and context-free derived characteristics to train a classifier. Hernault et al. [16] trained a discourse segmented using lexical and grammatical features using conditional random fields. Using lexico-syntactic, shallow syntactic, and contextual information, Joty et al. [17] used binary classifier for training to decide whether to set an EDU boundary for each.

(iii) Unsupervised approaches

Unsupervised learning in text segmentation involves training models without labeled datasets, relying on inherent patterns, structures, or similarities within the data to autonomously identify and segment text regions. Unlike supervised learning, there is no explicit guidance in the form of annotated ground truth during the training process.

Unsupervised approaches are based on lexical cohesion, which states that related terminology is more likely to be found in a coherent topic segment. TextTiling, the most prominent and early text segmentation method, was introduced by Hearst et al. [18].

Unsupervised learning in text segmentation is valuable in scenarios where obtaining labeled data is impractical or expensive. It has been applied in various domains, including content-based image retrieval, document clustering, and unsupervised document summarization, where discovering latent structures without explicit annotations is essential.

(iv) Non-linear character segmentation

Non-linear character segmentation refers to the process of identifying and isolating individual characters within a text image where the arrangement of characters may not follow a linear or sequential pattern. In contrast to linear character segmentation, where characters are arranged in a straightforward sequence, non-linear character segmentation is applied to cases where characters may be positioned in irregular or non-sequential patterns. This segmentation technique is particularly relevant in scenarios where text elements exhibit variations in alignment, orientation, or non-uniform spacing, different shapes and sizes [19]. [19] presented work on Character recognition of different scripts such as Roman, Devanagari, Oriya, Bangla and Japanese-Katakana. This work is based on non-linear multi-projection profiles measure.

[20] Tse et. al. developed OCR-independent character segmentation on grayscale document images using shortest-path. The shortest path strategy is used to find the correct segmentation path.

(v) Linear character segmentation

Linear character segmentation refers to the process of identifying and isolating individual characters within a text image where the arrangement of characters follows a linear or sequential pattern. In linear character segmentation, the characters are typically aligned in a straightforward sequence, making it relatively easier to extract each character individually.

Linear segmentation uses simple vertical segmentation. Here characters are deprived of their discriminative parts. However, this method increases the misclassification rate at a later stage [21]

(vi) Implicit segmentation

The implicit segmentation methods are known as recognition-based segmentation. These methods perform two tasks-character segmentation and recognition simultaneously. The primary advantage of implicit segmentation is that it eliminates the difficulty of segmentation. They do not necessitate any complicated "dissection" methods, and recognition mistakes are primarily caused by incorrect classifications. This type of segmentation is often built with rules that try to identify all of the character's segmentation points.

If fewer segments are used then it will reduce computation time. If there is overlapping between adjacent characters, then the problem will increase because it is necessary to recognize all possible combinations of valid characters rather than just valid characters.

Koteswara and Negi [22] proposed an HMM-based Telugu printed text recognition model. On the training set of word pictures, this model calculates statistical feature intensity and derivative of intensity using the sliding window technique.

In the Nasta'liq writing style, Naz et al. [23] proposed implicit segmentation of printed Urdu text lines. They employed Recurrent Neural Networks with Multidimensional Long Short-Term Memory (MDLSTM). The researcher generally examines a set of heuristics and information from the background [24,25], the foreground [26,27], or a mix of these [28,29] to generate segmentation cuts.

(vii) Projection-based segmentation

Projection-based segmentation is a technique used in image processing and computer vision to separate objects or regions of interest within an image. This method relies on analysing projections along one or more directions to identify distinctive features and boundaries within the image.

The basic idea behind projection-based segmentation involves projecting the pixel values of an image onto one or more axes and examining the resulting profiles. These profiles, known as projections, provide information about the distribution of pixel intensities along the selected directions. By analysing these projections, it becomes possible to detect variations in intensity that correspond to transitions between different objects or regions in the image.

(viii) Skeletonization-based based segmentation

skeletonization-based segmentation[30,31,32] leverages morphological operations to extract the essential structure of objects in an image, emphasizing their geometric features. This technique is valuable in scenarios where a simplified representation of object shapes is sufficient for subsequent analysis or where the focus is on identifying and segmenting elongated or linear structures. It recognizes character or word elements based on features like curvatures, boundaries, skeletons, angles, etc. that define region shape information.

(ix) Template Matching based segmentation

Template matching-based text segmentation is commonly used in document analysis, OCR (Optical Character Recognition), and information retrieval systems. It is effective when dealing with images containing structured or patterned text, such as scanned documents, forms, or images with consistent layouts. Template matching may face challenges in scenarios where the text varies significantly in size, orientation, or font style. Additionally, it may be sensitive to variations in illumination and background noise, requiring careful preprocessing to enhance robustness.

[33,34] used the sliding window and predefined character templates to determine the character's likely cutting spots. It begins with establishing the baseline. It then searches for matches between the templates and the text-image by swiping the templates over the text [33, 35, 36].

(x) Holistic Approaches in text segmentation

It involves considering the entire content and layout of an image as a unified entity rather than focusing on individual characters or components. Instead of breaking down the text into separate units, a holistic approach aims to understand and segment the text within the broader context of the entire document or image. This approach is known as the segmentation-free approach [37].

(xi) Region-Based Text segmentation

Region-based text segmentation divides an image into discrete areas based on specific criteria and then identifies and extracts text regions within these segments. This method focuses on the inherent qualities of regions rather than individual pixels, with the goal of grouping pixels with comparable characteristics together. When text sections exhibit specific visual qualities that may be recorded using feature extraction and clustering, region-based text segmentation is valuable.

Region-based text segmentation is advantageous in scenarios where the text regions exhibit specific visual properties that can be effectively captured through feature extraction and clustering

(xii) Contour Tracing based text Segmentation

Segmentation can be done by tracing the outer contour of a word [38]. Text segmentation methods based on contour tracing involve the use of contour detection and tracking techniques to identify and delineate the boundaries of text regions within an image.

Contour tracing methods evaluate the structural shape of characters as they are scanned, avoiding the problems created by thinning. Contour provides a clear outline of the characters, which might help in the under segmentation issues caused by overlapping characters [39].

Contour tracing is particularly useful in scenarios where text regions have distinct and well-defined contours, making it feasible to trace and extract them from the background.

4. Applications of Image Text Segmentation across Various Domains

Image text segmentation has numerous applications across various domains due to its capability to isolate and extract textual information from images. Some notable application areas include:

(i) Document Analysis

In document analysis, text segmentation is crucial for extracting text from scanned documents, handwritten notes, or printed pages. It facilitates tasks such as OCR (Optical Character Recognition), document classification, and information retrieval.

(ii) Automated Captioning and Image Tagging

Automatic text detection from images is useful for accessing and utilising textual data [9]. The text in the image has importance information. Image text segmentation is used to identify and extract textual content from images, enabling automated captioning and tagging. This is particularly valuable in applications like social media, where images are annotated with relevant text for improved accessibility and searchability.

(iii) Visual Search Engines

Text segmentation enhances visual search engines by isolating and indexing text within images. This improves the accuracy and relevance of search results, making it easier to find images containing specific textual content.

(iv) Medical Imaging

In medical imaging, text segmentation is employed to extract and analyse textual information from radiological images, pathology reports, and medical documents. This aids in automating the extraction of critical information for diagnostics and patient records.

(v) News and Media Analysis

Image text segmentation is utilized in news and media analysis to extract text from images, facilitating automated processing for categorization, sentiment analysis, and content understanding.

Conclusion

This comprehensive comparative study on text segmentation techniques, aimed to provide valuable insights into the strengths, weaknesses, and applicability of each technique. The comparative analysis also emphasized the importance of considering specific application requirements when choosing a text segmentation technique.

As the field of natural language processing continues to evolve, there are opportunities for further research and innovation in text segmentation. Future work could explore hybrid approaches that combine the strengths of rule-based, statistical, and machine learning techniques to achieve improved adaptability and robustness. Additionally, the development of techniques capable of handling low-resource languages and domain-specific jargon remains an avenue for exploration.

As the journey in natural language processing progresses, this research contributes to understanding and the way for advancements in text segmentation, ultimately enhancing the efficiency and efficacy of a wide array of NLP applications.

Reference

[1]. Liu, Chuanhan, Yongcheng Wang, and Fei Zheng. "Automatic text summarization for dialogue style." In 2006 IEEE International Conference on Information Acquisition, pp. 274-278. IEEE, 2006.

[2]. Osman, Deanna J., and John L. Yearwood. "Opinion search in web logs." In Proceedings of the eighteenth conference on Australasian database-Volume 63, pp. 133-139. 2007.

[3]. Wu, Yun, Yan Zhang, Si-ming Luo, and Xiao-jie Wang. "Comprehensive information based semantic orientation identification." In 2007 International Conference on Natural Language Processing and Knowledge Engineering, pp. 274-279. IEEE, 2007.

[4]. Gao, Yang, Li Zhou, Yong Zhang, Chunxiao Xing, Yigang Sun, and Xianzhong Zhu. "Sentiment classification for stock news." In 5th International Conference on Pervasive Computing and Applications, pp. 99-104. IEEE, 2010.

[5]. Xia, Huosong, Min Tao, and Yi Wang. "Sentiment text classification of customers reviews on the Web based on SVM." In 2010 Sixth International Conference on Natural Computation, vol. 7, pp. 3633-3637. IEEE, 2010.

[6]. Potrus, Moayad Yousif, Umi Kalthum Ngah, and Bestoun S. Ahmed. "An evolutionary harmony search algorithm with dominant point detection for recognition-based segmentation of online Arabic text recognition." Ain Shams Engineering Journal 5, no. 4 (2014): 1129-1139.

[7]. Brants, Thorsten, Francine Chen, and Ioannis Tsochantaridis. "Topic-based document segmentation with probabilistic latent semantic analysis." In Proceedings of the eleventh international conference on Information and knowledge management, pp. 211-218. 2002.

[8]. Flejter, Dominik, Karol Wieloch, and Witold Abramowicz. "Unsupervised methods of topical text segmentation for polish." In Proceedings of the Workshop on Balto-Slavonic Natural Language Processing, pp. 51-58. 2007.

[9]. J. M. Patel and A. A. Desai, "Gujarati Text Localization, Extraction and Binarization from Images", International Journal of Computer Sciences and Engineering vol. 6, no. 8 (2018),pp. 714-724.

[10]. Li, Jing, Billy Chiu, Shuo Shang, and Ling Shao. "Neural text segmentation and its application to sentiment analysis." IEEE Transactions on Knowledge and Data Engineering 34, no. 2 (2020): 828-842.

[11]. Nafiz Arica and Fatos T. Yarman-Vural,"An Overview of Character Recognition Focused on Off-Line Handwriting" in IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS, May 2001.

[12]. D. Brodic and Z. Milivojevic,"A New Approach to Water Flow Algorithm for Text Line Segmentation" in Journal of Universal Computer Science, vol. 17, no. 1,2011.

[13]. R. S. Kunte and R. D. Sudhaker Samuel, A simple and efficient optical character recognition system for basic symbols in printed Kannada text: Sadhana, 32, 2007, 521–533.

[14]. Z. Razak, K. Zulkiflee, R. Salleh, M. Yaacob and E. Mohd, Tamil: A real-time line segmentation algorithm for an offline overlapped handwritten jawi character recognition chip, Malaysian Journal of Computer Science, 20, 2007, 171–182.

[15]. S. Fisher and B. Roark, “The utility of parse-derived features for automatic discourse segmentation,” in ACL, vol. 45, no. 1, 2007, p. 488.

[16]. H. Hernault, D. Bollegala, and M. Ishizuka, “A sequential model for discourse segmentation.” in CICLing, 2010, pp. 315–326.

[17]. S. Joty, G. Carenini, and R. T. Ng, “Codra: A novel discriminative framework for rhetorical analysis,” Comput. Linguist., vol. 41, pp. 385–435, 2015.

[18]. A. Hearst, “Texttiling: Segmenting text into multi-paragraph subtopic passages,” Comput. Linguist., vol. 23, no. 1, pp. 33–64, 1997

[19]. Santosh, K. C., and Laurent Wendling. "Character recognition based on non-linear multi-projection profiles measure." Frontiers of Computer Science 9 (2015): 678-690.

[20]. J. Tse, C. Jones, D. Curtis and E. Yfantis, "An OCR-independent character segmentation using shortest-path in grayscale document images," Sixth International Conference on Machine Learning and Applications (ICMLA 2007), Cincinnati, OH, USA, 2007, pp. 142-147

[21]. Jayarathna, U. K. S., and G. E. M. D. C. Bandara. "New segmentation algorithm for offline handwritten connected character segmentation." In First International Conference on Industrial and Information Systems, pp. 540-546. IEEE, 2006.

[22]. Rao D. Koteswara and Atul Negi, "An implicit segmentation approach for Telugu text recognition based on hidden Markov models.", In Advances in Signal Processing and Intelligent Recognition Systems, (2016), pp. 633-644.

[23]. Naz Saeeda, Arif Iqbal Umar, Riaz Ahmed, Muhammad Imran Razzak, Sheikh Faisal Rashid and Faisal Shafait, "Urdu Nasta’liq text recognition using implicit segmentation based on multi-dimensional long short term memory neural networks", SpringerPlus, vol. 5, no. 1, (2016), pp. 1-16.

[24]. M. Cheriet, Y.S. Huang and C.Y. Suen, “Background region based algorithm for the segmentation of connected digits”, In Proceedings of the 11th International Conference on Pattern Recognition, (1992), pp. 619–622.

[25]. Lu, Z., Chi, Z., Siu, W., Shi, P., “A background-thinning-based approach for separating and recognizing connected handwritten digit strings”, Pattern Recognit, vol. 32, (1999), pp. 921–933.

[26]. E. Lethelier, M. Leroux and M. Gilloux, “An automatic reading system for handwritten numeral amounts on french checks”, In Proceedings of 3rd International Conference on Document Analysis and Recognition, (1995), pp. 92–97.

[27]. J. Sadri, C. Y. Suen and T. D. Bui, “A genetic framework using contextual knowledge for segmentation and recognition of handwritten numeral strings”, Pattern Recognit, vol. 40, (2007), pp. 898–919.

[28]. Y. K.Chen and J. F. Wang, “Segmentation of single- or multiple-touching handwritten numeral string using background and foreground analysis”, IEEE Trans. Pattern Anal. Mach. Intell, vol. 22, no. 11, (2000), pp. 1304–1317.

[29]. L. S. Oliveira, R. Sabourin, F. Bortolozzi, and C. Y. Suen, “Automatic recognition of handwritten numerical strings: a recognition and verification strategy”, IEEE Trans. Pattern Anal. Mach. Intell, vol. 24, no. 11, (2002), pp. 1438–1454.

[30]. F. U. Qomariyah, and W. F. Mahmudy, “The segmentation of printed arabic characters based on interest point”, Journal of Telecommunication, Electronic and Computer Engineering, vol. 9, (2017), pp. 19–24.

[31]. M. M. Altuwaijri and M. A. Bayoumi, “A thinning algorithm for arabic characters using art2 neural network”, IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing, vol. 45, no. 2, (1998), pp. 260–264.

[32]. B. F. H. Timsari, “Morphological approach to character recognition in machine printed persian words”, In Proceeding of SPIE. Document Recognition III, (1996).

[33]. R. Saabni, “Efficient recognition of machine printed arabic text using partial segmentation and hausdorff distance”, In 2014 6th International Conference of Soft Computing and Pattern Recognition (SoCPaR), (2014), pp. 284–289.

[34]. Y. M. Alginahi, “A survey on arabic character segmentation”, Int. J. Document Anal. Recogn, vol. 16, no. 2, 2013, pp. 105–126.

[35]. A. Lawgali, “A survey on arabic character recognition”, International Journal of Signal Processing, Image Processing and Pattern Recognition, vol. 8, no. 2, (2015), pp. 401–426.

[36]. Y. Zhang, Z. Q. Zha and L. F. Bai, “A license plate character segmentation method based on character contour and template matching”, In Applied Mechanics and Materials, vol. 333, (2013), pp. 974–979.

[37]. J. Ahmad, “Optical character recognition system for arabic text using cursive multi-directional approach”, Journal of Computer Science, vol. 3, (2007), pp. 549–555.

[38]. M. Omidyeganeh, K. Nayebi, R. Azmi, and A. Javadtalab, “A new segmentation technique for multi font farsi/arabic texts”, In Proceedings. (ICASSP’05). IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 2. (2005).

[39]. R. Saabni, “Efficient recognition of machine printed arabic text using partial segmentation and hausdorff distance”, In 2014 6th International Conference of Soft Computing and Pattern Recognition (SoCPaR), (2014), pp. 284–289.