Methods and Development of Chinese Word Tokenization

Zhenghan Fang

School of Electronic Engineering & Computer Science, Queen Mary University of London, London, UK

ec24118@qmul.ac.edu

Abstract. Chinese word tokenization is an important task in natural language processing and has undergone significant evolution with artificial intelligence. This paper reviews the historical progression and contemporary methodologies for Chinese word segmentation. The paper examines the traditional character-based approaches, which rely on dictionaries and pattern matching, and transition into machine learning-based techniques that utilize statistical models and neural networks. A particular focus is given to the recent developments in deep learning, including the application of recurrent neural networks (RNNs), long short-term memory networks (LSTMs), and transformer models like BERT. The review also highlights innovative approaches such as memory networks and sub-character tokenization, which have shown promising results in improving segmentation accuracy and computational efficiency. Furthermore, the paper discusses the challenges faced in tokenization, such as handling out-of-vocabulary words and the integration of syntactic and semantic information. The paper concludes with insights on the future directions of Chinese word tokenization, emphasizing the potential of unsupervised learning and the need for more robust evaluation frameworks.

Keywords: Chinese word tokenization, natural language processing, machine learning, deep learning.

1. Introduction

The rapid expansion of digital communication and the increasing prevalence of Mandarin Chinese in global discourse have underscored the critical role of Chinese word tokenization in the field of Natural Language Processing (NLP). Unlike languages with explicit word delimiters, Chinese scripts present a continuous stream of characters that necessitate sophisticated tokenization techniques for effective language analysis and information extraction. The task of Chinese word tokenization is not merely a technical challenge but a gateway to higher-level NLP applications, including machine translation and sentiment analysis [1].

Tokenization in Chinese presents unique challenges compared to English due to the structural differences between the languages. In English, tokenization often relies on whitespace and punctuation to demarcate words, which is facilitated by the fact that English is a space-delimited language. Each word is separated by spaces, making it relatively straightforward to split text into tokens. In contrast, Chinese does not use spaces between words, and each character can carry its own meaning. This absence of explicit word boundaries means that tokenization must often rely on more complex algorithms to identify word-like sequences, known as terms, within the text. Chinese tokenization typically involves segmenting the text into the most meaningful units, which can be single characters, multi-character words, or phrases [2].

One of the key challenges in Chinese tokenization is the vast number of homophones and the context-dependent nature of word meanings. A single character can have multiple meanings depending on its usage, which requires sophisticated algorithms to disambiguate. Additionally, the granularity of tokenization in Chinese can significantly affect the performance of downstream NLP tasks. For instance, a sub-character level tokenization that considers radicals or phonetic components can provide more information but may complicate the process.

Also, ancient Chinese tokenization is an additional challenge. Due to the evolution of the language, it has many differences compared to Modern Standard Chinese. The script used in ancient texts, such as Classical Chinese, is more archaic and often employs a richer and more complex vocabulary than its modern counterpart. This can make it difficult for modern tokenization algorithms, which are typically trained on contemporary language data, to accurately segment the text. Additionally, ancient texts often do not use punctuation, which further complicates the process of identifying sentence and phrase structures [3]. Another challenge is the presence of variant characters and the use of classical Chinese, which includes a large number of obsolete characters that are no longer in common use. This can lead to a higher rate of unknown tokens if the tokenizer's vocabulary is not extended to include these characters. Moreover, the context and cultural references in ancient texts are often deeply rooted in historical events and philosophical concepts. An idiom or abbreviation for a historical event may not be well-captured by modern tokenization methods. This requires tokenizers to be aware of historical linguistic nuances and to handle classical Chinese in a way that preserves the semantic richness of the text.

This paper conducts a thorough review of the methods used for segmenting Chinese text into words, a critical task in natural language processing. It traces the evolution from traditional dictionary-based approaches to modern machine learning and deep learning techniques, with a spotlight on recent innovations like deep learning models and sub-character tokenization. The paper also addresses the unique challenges of Chinese tokenization, such as handling homophones and the complexities of ancient Chinese scripts. It discusses the potential of unsupervised learning and the necessity for robust evaluation methods, providing a comprehensive overview of the current state and future prospects of Chinese word tokenization.

2. Methods for Chinese Word Tokenization

2.1. Unigram wordpiece

Bidirectional Encoder Representations from Transformers (BERT) uses a specific type of tokenizer that is designed to handle the nuances of the language it is trained on. For English and other languages that use whitespace to separate words, BERT typically uses a WordPiece tokenizer [4].

Unigram WordPiece is a tokenization algorithm that segments text into subword units. It operates by considering the entire vocabulary as a set of tokens and then iteratively merges the most frequent word pairs into single tokens. This process continues until a desired vocabulary size is reached or a specific threshold of frequency is surpassed. Unigram WordPiece is particularly adept at handling out-of-vocabulary words by breaking them down into their constituent parts [5].

However, when it comes to Chinese tokenization, Unigram WordPiece faces several challenges. Over segmentation is one of the problems, Chinese characters often carry individual meanings, and Unigram WordPiece might treat them as independent tokens, especially if they are not frequently occurring together as a pair in the training data. Especially to idiomatic expressions, they may be split into individual characters by Unigram WordPiece, losing the idiomatic meaning.

In the context of Ancient Chinese tokenization, the vocabulary and syntax of Ancient Chinese differ significantly from modern language, and many words and phrases are not represented in contemporary corpora. Unigram WordPiece, which is heavily reliant on frequency, may fail to recognize archaic terms and phrases, leading to a high rate of tokenization errors. And due to the lack of historical materials, it is not practical to increase the corpora for Ancient Chinese.

2.2. Glyph-based and Pronunciation-based tokenization

Glyph-based and Pronunciation-based tokenization are two novel approaches to better capture the linguistic characteristics of the Chinese writing system. These methods, known as Sub-character tokenization, aim to leverage information below the character level, which traditional tokenization methods often overlook [6].

Glyph-based Tokenization encodes Chinese characters into sequences that represent their visual structure or radicals. For instance, characters can be broken down into semantically meaningful components using input methods like Wubi. This encoding captures the semantic richness inherent in character shapes, as characters with common radicals often share related meanings. By encoding characters into sequences of these radicals, glyph-based tokenization embeds semantic information directly into the tokenization process. For instance, characters that share common radicals, such as '氵' (water), are likely related in meaning. The advantage of glyph-based tokenization is its ability to embed semantic information directly into the tokenization process, potentially enhancing the model's understanding of the text. However, it may not fully capture the pronunciation aspect, which is crucial for languages like Chinese with a complex grapheme-phoneme relationship [7].

Pronunciation-based Tokenization, on the other hand, focuses on the phonetic properties of characters. It uses transliterations such as Pinyin or Zhuyin to represent characters phonetically. This method is advantageous because it directly addresses the pronunciation of characters, which is essential for understanding the language's sound patterns and handling homophones effectively. For example, pronunciation-based Sub-character tokenizers can encode homophones into the same transliteration sequences, making the model robust against homophone typos. In tasks such as conversation, or machine translation, distinguishing between homophones is essential for accurate disambiguation. Also, in the process of language development many homophones in ancient Chinese are merged into one word, or one word in ancient Chinese is broken down into multiple homophones with different meanings [8].

The integration of "Glyph-based Tokenization" and "Pronunciation-based Tokenization" offers significant advantages when applied to the task of tokenizing Classical Chinese texts. these methods are adept at overcoming the limitations posed by a scarcity of corpus data, as they do not rely heavily on the frequency of occurrence of rare characters. This feature is particularly beneficial for Classical Chinese, where many characters are not only infrequently used but also challenging to digitize due to their complexity.

Also, by aligning with the intrinsic input habits of the Chinese language, these tokenization strategies can inherently correct textual errors to a certain extent, thereby enhancing the robustness of the NLP models. This is crucial for the accurate processing of ancient texts, which often contain characters and phrases that have evolved significantly or become obsolete in modern usage.

However, the same limitation of insufficient corpus data can also be a drawback. The understanding and interpretation of Classical Chinese texts may require human intervention to a certain degree. This is because the pronunciation and writing of modern characters are probably completely different from that of ancient Chinese, or there are errors in the writing records process. This was very common because the ancient texts had to be copied by hand. So, expert guidance is necessary to ensure the accuracy and depth of analysis.

2.3. Joint Chinese Word Segmentation and Part-of-Speech Tagging via Two-way Attentions of Auto-analyzed Knowledge

The study on joint Chinese Word Segmentation (CWS) and Part-of-Speech (POS) tagging, which are fundamental tasks in Chinese language processing. This method uses a neural model named TWASP, which employs a two-way attention mechanism to integrate contextual features and their corresponding syntactic knowledge for each input character. It is designed to enhance the performance of CWS and

POS tagging by leveraging auto-analyzed syntactic knowledge generated by existing Natural Language Processing toolkits [9].

The TWASP model is anchored on a character-based sequence labeling paradigm, which is a departure from the more conventional word-based or sentence-based models. This choice is motivated by the unique characteristics of the Chinese language, where words are not explicitly segmented by spaces, thus necessitating a more granular approach to both segmentation and POS tagging. The model's input is a sequence of characters, denoted as X , and the output is a corresponding sequence of joint labels, represented as Y . Each y is a composite tag that includes both the segmentation and POS information for the character x.The two-way attention mechanism of TWASP, which is designed to incorporate contextual features and their associated syntactic knowledge for each character in the input sequence. This mechanism is distinct in its ability to process features and knowledge separately, allowing the model to discern and prioritize the most relevant information for accurate tagging. Contextual features and knowledge instances are extracted using auto-analyzed results from existing NLP toolkits, such as the Stanford CoreNLP Toolkit and the Berkeley Neural Parser. These toolkits provide a rich source of syntactic knowledge, including POS labels, syntactic constituents, and dependency relations, which are crucial for understanding the structure and semantics of the text.

The model's architecture is composed of two main components: the backbone model for joint CWS and POS tagging, and the two-way attention module. The backbone model processes the input character sequence and generates an initial set of embeddings, which serve as the basis for the attention mechanism. It operates in two distinct pathways, referred to as the feature way and the knowledge way. In the feature way, the model attends to the contextual features associated with each character, while in the knowledge way, it attends to the corresponding syntactic knowledge instances. The attention weights are computed separately for each pathway, and the resulting attention vectors are concatenated to form a comprehensive representation that guides the tagging process.

for each character x_i, the model computes the attention weights $a_{s,i,j}$, for each context feature $s_{i,j}$ in s_i using a feed-forward attention module. The weights are normalized using a softmax function, ensuring that the model focuses on the most informative features. Similarly, the model computes attention weights $a_{k,i,j}$ for each knowledge instance $k_{i,j}$ in K_i. The final attention vectors for each character are obtained by concatenating the attention vectors from both pathways: $a_i = a_{s,i} \oplus a_{k,i}$ .The joint tagging process is facilitated by the combined attention vectors, which are used to refine the embeddings generated by the backbone model. For each character x_i, the model concatenates the character's vector h_i from the backbone model with the corresponding attention vector a_i. This concatenated vector is then passed through a fully connected layer to align the dimensions for final prediction.

The model employs Conditional Random Fields (CRF) to estimate the probability of joint labels, capturing the sequential dependencies and ensuring consistent labeling. The CRF layer considers the output from the fully connected layer and the preceding tag to determine the likelihood of the current tag. The model's efficacy is substantiated through comprehensive testing on five benchmark datasets, with results indicating that TWASP surpasses standard models, achieving cutting-edge performance in joint CWS and POS tagging. The two-way attention mechanism proves to be highly effective in utilizing auto-analyzed syntactic knowledge, even when its accuracy is less than perfect.

2.4. Wordhood Memory Networks

The study from Wordhood Memory Networks introduces WMSEG, a groundbreaking neural framework that employs wordhood memory networks to significantly enhance the task of CWS. This model represents a paradigm shift in the way contextual features, specifically wordhood information, are integrated into neural networks for CWS. Traditionally, wordhood information has been a valuable component in character-based segmenters, but it has been somewhat overlooked in recent neural models [10].

WMSEG is a memory module that functions as a key-value store for n-grams and their wordhood scores. This module allows the model to access rich contextual information that is often crucial for disambiguating word boundaries in Chinese, a language lacking explicit word delimiters. The framework's approach to constructing the lexicon, which is simply a list of n-grams, is both innovative and pragmatic. It harnesses unsupervised wordhood measures such as Accessor Variety, Pointwise Mutual Information, and Description Length Gain to identify potential word boundaries directly from the text. This unsupervised approach stands in contrast to traditional methods that rely heavily on annotated datasets or predefined lexicons. For every character, the module considers all n-grams that contain it and retrieves the corresponding wordhood information to produce an enhanced output vector. This vector is instrumental in aiding the decoder to assign the correct segmentation label to each character and use different encoders allows WMSEG to be adaptable to various scenarios and data characteristics, showcasing its flexibility.

WMSEG achieving state-of-the-art performance across five benchmark datasets. This success can be attributed to the model's ability to effectively capture and utilize wordhood information. The memory mechanism models wordhood information in a way that is intuitive and aligned with the strengths of neural networks, leading to significant improvements in segmentation accuracy. The robustness of WMSEG is further underscored by its consistent performance across different wordhood measures and its resilience in cross-domain experiments, which is a common challenge for many NLP models. And one of the most compelling aspects of WMSEG is its ability to handle out-of-vocabulary (OOV) words with greater accuracy than previous models. This is particularly important for CWS, as OOV words are a frequent source of errors in language processing tasks. The model's memory module provides it with the contextual awareness needed to make informed predictions about word boundaries, even when confronted with new or unseen words.

The advantages of WMSEG include its state-of-the-art performance, demonstrated through superior results on multiple benchmark datasets. Its flexibility is highlighted by the ability to combine the memory module with different encoder-decoder pairs, making the framework highly adaptable. The model also shows robustness across various domains and is not limited to specific wordhood measures, indicating its effectiveness in diverse settings. Furthermore, WMSEG does not rely on external resources, as it utilizes unsupervised measures for lexicon construction, making it more accessible. It also shows improved handling of OOV words, a common challenge in word segmentation.

However, the framework has some disadvantages. The incorporation of memory networks adds complexity to the model architecture, potentially increasing the computational cost. The performance might be sensitive to the initialization of the memory keys and values, although the paper notes that random initialization works well in their experiments. A deeper analysis on how the memory module learns from wordhood information is lacking, and while the effectiveness of the memory module is shown, more insight into its learning process would be beneficial. The effectiveness relies on the quality of the lexicon constructed from wordhood measures; therefore, poor lexicon construction could degrade performance. Additionally, it is unclear how well the approach would generalize to other languages without modifications.

3. Conclusion

In conclusion, the field of Chinese word tokenization has witnessed remarkable advancements, transitioning from traditional dictionary-based methods to sophisticated machine learning and deep learning models. The evolution of tokenization techniques has been instrumental in enhancing the performance of various NLP tasks, such as machine translation and sentiment analysis. This paper has provided a comprehensive review of the historical progression and contemporary methodologies for Chinese word segmentation, highlighting the pivotal role of deep learning in this domain.

The exploration of character-based, pattern matching, and machine learning-based techniques has laid the groundwork for more advanced models like RNNs, LSTMs, and transformer models such as BERT. These models have demonstrated significant improvements in segmentation accuracy and computational efficiency. Innovative approaches like memory networks and sub-character tokenization have shown promise in addressing the challenges of out-of-vocabulary words and the integration of syntactic and semantic information.

Looking ahead, the future of Chinese word tokenization is poised to be influenced by the potential of unsupervised learning and the development of more robust evaluation frameworks. Unsupervised learning, in particular, holds the promise of reducing reliance on annotated datasets, thus making the tokenization process more accessible and adaptable to various domains. The creation of comprehensive evaluation frameworks will facilitate the systematic assessment of tokenization models, ensuring their reliability and effectiveness across different applications. And it is also anticipated that interdisciplinary research will play a critical role in driving innovation in this field. The integration of linguistics, computer science, and artificial intelligence will likely yield novel insights and methodologies that can further refine tokenization techniques. Additionally, the increasing availability of computational resources and the development of more efficient algorithms will enable the processing of large-scale datasets, leading to more accurate models.

In summary, the progress in Chinese word tokenization is a testament to the dynamic interplay between linguistic theory, technological innovation, and practical application. As research continues to push the boundaries of what is possible, the future of Chinese word tokenization appears bright, with the potential to unlock new frontiers in language processing and understanding.

References

[1] Huang, C. R., Šimon, P., Hsieh, S. K., & Prévot, L. (2007, June). Rethinking chinese word segmentation: tokenization, character classification, or wordbreak identification. In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions (pp. 69-72).

[2] Wang, D., Li, Y., Jiang, J., Ding, Z., Jiang, G., Liang, J., & Yang, D. (2024). Tokenization Matters! Degrading Large Language Models through Challenging Their Tokenization. arXiv preprint arXiv:2405.17067.

[3] Blouin, B., Huang, H. H., Henriot, C., & Armand, C. (2023, December). Unlocking transitional Chinese: word segmentation in modern historical texts. In Proceedings of the Joint 3rd International Conference on Natural Language Processing for Digital Humanities and 8th International Workshop on Computational Linguistics for Uralic Languages (pp. 92-101).

[4] Kolonin, A., & Ramesh, V. (2022). Unsupervised tokenization learning. arXiv preprint arXiv:2205.11443.

[5] Song, X., Salcianu, A., Song, Y., Dopson, D., & Zhou, D. (2020). Fast wordpiece tokenization. arXiv preprint arXiv:2012.15524.

[6] Si, C., Zhang, Z., Chen, Y., Qi, F., Wang, X., Liu, Z., ... & Sun, M. (2023). Sub-character tokenization for Chinese pretrained language models. Transactions of the Association for Computational Linguistics, 11, 469-487.

[7] Yan, H., Qiu, X., & Huang, X. (2020). A graph-based model for joint chinese word segmentation and dependency parsing. Transactions of the Association for Computational Linguistics, 8, 78-92.

[8] Zhang, Y., & Clark, S. (2007, June). Chinese segmentation with a word-based perceptron algorithm. In Proceedings of the 45th annual meeting of the association of computational linguistics (pp. 840-847).

[9] Tian, Y., Song, Y., Ao, X., Xia, F., Quan, X., Zhang, T., & Wang, Y. (2020, July). Joint Chinese word segmentation and part-of-speech tagging via two-way attentions of auto-analyzed knowledge. In Proceedings of the 58th annual meeting of the association for computational linguistics (pp. 8286-8296).

[10] Tian, Y., Song, Y., Xia, F., Zhang, T., & Wang, Y. (2020, July). Improving Chinese word segmentation with wordhood memory networks. In Proceedings of the 58th annual meeting of the association for computational linguistics (pp. 8274-8285).