Context-Aware Word Segmentation for Chinese Real-World Discourse

Kaiyu Huang, Junpeng Liu, Jingxiang Cao and Degen Huang*
School of Computer Science, Dalian University of Technology
{kaiyuhuang,liujunpeng_nlp}@mail.dlut.edu.cn
{caojx,huangdg}@dlut.edu.cn

Abstract

Previous neural approaches achieve significant progress for Chinese word segmentation (CWS) as a sentence-level task, but it suffers from limitations on real-world scenario. In this paper, we address this issue with a context-aware method and optimize the solution at document-level. This paper proposes a three-step strategy to improve the performance for discourse CWS. First, the method utilizes an auxiliary segmenter to remedy the limitation on pre-segmenter. Then the context-aware algorithm computes the confidence of each split. The maximum probability path is reconstructed via this algorithm. Besides, in order to evaluate the performance in discourse, we build a new benchmark consisting of the latest news and Chinese medical articles. Extensive experiments on this benchmark show that our proposed method achieves a competitive performance on a document-level real-world scenario for CWS.

1 Introduction

Downstream tasks in Chinese natural language processing (NLP) leverage word-level information to construct architectures. In recent years, some technologies gradually replace the word-level information in order to alleviate segmentation errors and word sparse problems (Li et al., 2019). However, it may lose all of the other word-level information (e.g., Part-of-speech, and Dependency parsing). Chinese word segmentation (CWS) is still essential for downstream Chinese NLP tasks.

Xue (2003) formalized CWS task as a sequence labeling problem. The performance for CWS has achieved significant progress via statistical machine learning (Zhao and Kit, 2008; Zhao et al., 2010) and neural network methods (Cai et al., 2017; Zhou et al., 2017; Yang et al., 2017; Ma et al., 2018). In particular, recent years have also seen a new supervised learning paradigm in applying BERT or other pre-training models for sequence labeling problems (Huang et al., 2019; Meng et al., 2019; Tian et al., 2020). Various fine-tuning methods can improve the performance for in-domain CWS significantly and easily. Previous researches almost perform like humans with a nearly 2% error rate via pre-training methods. And neural methods do not rely on the hand-craft feature engineering, compared with statistical machine learning methods. However, recent state-of-the-art neural methods always suffer from two limitations as follow:

1) Effective neural network methods and fine-tuning methods based on pre-training models need large annotated corpora to train. The performance will take a nosedive under a low-resource scenario.

2) Neural network methods are good at processing short sentences instead of long sentences or document-level texts, especially for Chinese word segmentation. Although some neural architectures try to alleviate this problem, e.g., long-short term memory (LSTM) neural network, the performance still drops dramatically under a long maximum length of sentences.

Both two limitations are reflected under real-world document-level texts apparently. Many Chinese tasks are discourse processing researches in most cases, e.g., text classification and machine translation. So the effectiveness for real-world document-level CWS affects the performance on these tasks. Real-world texts are always time-efficient. Not only cross-domain but also time-validity leads to the issue of out-of-vocabulary (OOV) words. Even though previous researches alleviate the issues for cross-domain CWS, the issue for time-efficient CWS is hard to solve with similar inspiration by these effective methods. Most methods start with data-driven to improve the performance for cross-

*Corresponding author

Proceedings of the 2nd International Workshop on Discourse Processing, pages 22–28

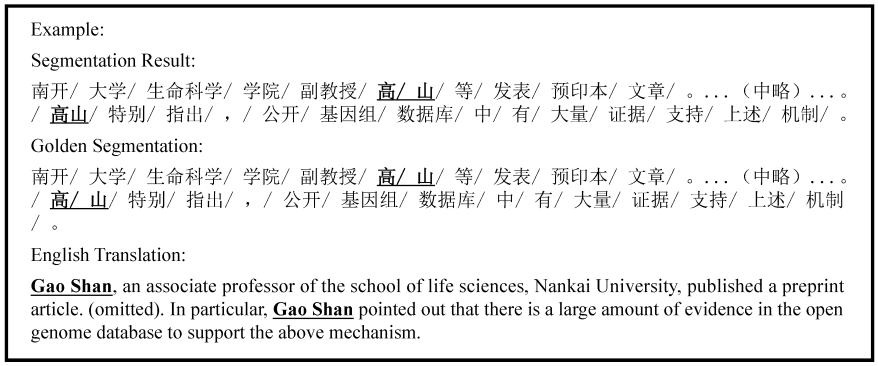

Figure 1: The limitation on the document-level Chinese word segmentation

domain CWS. For instance, previous works incorporate the domain dictionary and pre-training embedding into neural network methods (Zhang et al., 2018; Zhao et al., 2018; Ye et al., 2019). Annotating a small number of training corpora is the most effective and simple method. However, large annotated corpus cannot be updated immediately to deal with the latest news and dialogues, due to the high cost of the hand-craft annotated work. Similarly, training an effective pre-training model is time-consuming. It is impractical to update the model and annotated corpus with the latest data.

Furthermore, previous effective methods always suffer from the weakness of robustness for discourse CWS, as shown in Figure 1. The word "Gao Shan" occurs twice with the context of similar semantics in the discourse. However, it is segmented into different splits. Because of the limitation on the maximum length of input sequences, segmentation consistency is not guaranteed.

Previous works for in-domain and cross-domain CWS are always character-level and sentence-level. Li and Xue (2014) proposes an effective method for patent domain CWS via integrating the document-level features, but it is still a sentence-level optimization . Yan et al. (2017) utilizes multiple constraint rules to alleviate the issues on specific domains. This paper proposes a context-aware unsupervised method to alleviate the above issues for Chinese word segmentation, instead of adopting multiple constraint rules. And it is not limited to a specific domain. The method is aware of a global receptive field in the entire discourse. It utilizes the document-level information directly to improve the performance for CWS on the real-world discourse scenario. In particular, the words that recur in discourse are rejudged by our proposed method. The uncertain words are also reconsidered. The method consists of three steps: a word-lattice based pre-segmenter, a rejudged module, and a context-aware algorithm. First, the pre-segmenter achieves high performance on in-vocabulary words. Then the rejudged module chooses the uncertain splits as potential out-of-vocabulary words. Finally, the core context-aware algorithm utilizes the document-level information to screen the uncertain splits. The sequence labeling task is to find the maximum probability path. A new path is reconstructed through the three steps. To evaluate our method, we build a new benchmark with document-level texts for CWS. It contains the latest news and Chinese medical articles. Extensive experiments show that our proposed method is effective in discourse area and achieves a competitive performance for real-world CWS.

To sum up, our main contributions are threefold: 1) To the best of our knowledge, our proposed method is the first work to adopt a document-level unsupervised learning algorithm for CWS in a real-world scenario. It only takes the information about the current discourse itself. The benefit is that there is no need to do maintenance work for external resources constantly. 2) The method acts on the global field of discourse. It can alleviate the issue of segmentation inconsistency effectively. 3) We propose a new benchmark to evaluate the performance for CWS on real-world discourse scenario.[1]

2 Methodology

Figure 2 shows the entire process of our proposed method. It consists of three steps. In the first two steps, we utilize a pre-segmenter and an auxiliary segmenter to segment the sentences. The two segmenters generate two segmentation results R

[1]Our code are available at https://github.com/koukaiu/dlut-nihao

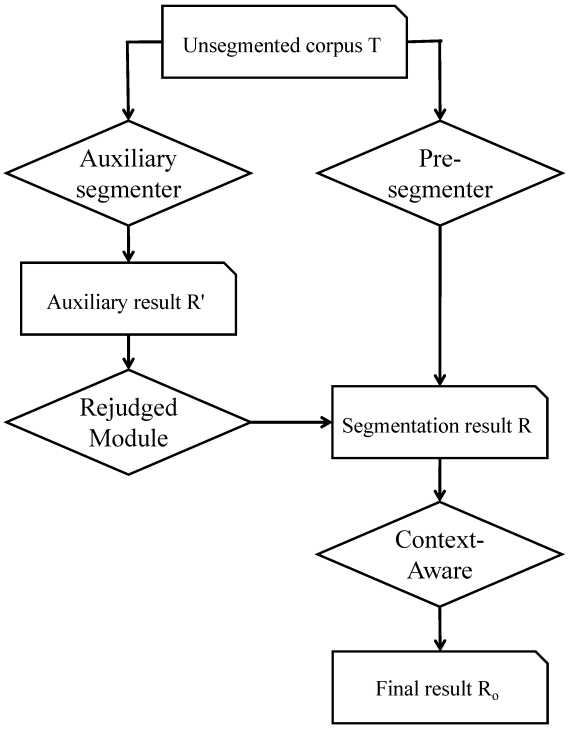

Figure 2: The process of the context-aware method

and R' respectively. It can distribute segmentation splits with different reliability. The pre-segmenter and auxiliary segmenter are all sentence-level CWS methods. In the third step, the context-aware algorithm leverages the document-level information to determine the final boundary of these splits. It is global optimization to revise the results in the prior steps. The final segmentation result R_o is obtained by this optimization.

2.1 Pre-segmenter

The word-lattice based method is studied for a long time and it is gradually replaced by neural methods. However, the method retains a high performance on in-vocabulary words (Kudo et al., 2004). It depends on the robustness of the word lattice. Given an unsegmented sentence S, the method builds an undirected graph with a system lattice. The graph consists of nodes and edges respectively. Every node represents the splits which could be a word, and the edges are paths with the transition probability. CWS is transferred into a task of searching the maximum probability path of the undirected graph. The word-lattice based model is essentially a statistical machine learning method, and the transition probability is trained with a hand-craft feature template inevitably. To simplify this process, we adopt a simple and base feature template that consists of the current word itself and the sliding window with a front and back distance of two. This process does not need much knowledge. The effort and cost are similar to those of building embedding layer in neural methods.

2.2 Auxiliary Segmenter

The pre-segmenter can provide the segmentation result of high accuracy on in-vocabulary words. However, the recognition capability of out-of-vocabulary words is weak by the pre-segmenter. We leverage an auxiliary segmenter to rejudge uncertain splits in the first step for the potential to become words. In order to capture the flexible lexical features of characters, the auxiliary segmenter employs a BERT fine-tuning learning paradigm on the character-level CWS. Given the same unsegmented sentence S in the first step, the character sequence of the sentence maps onto corresponding

Algorithm 1 The document-level context-aware optimization

Input:

The pre-segmentation result of characters with labels $R = (t_1, t_2, t_3 \cdots t_n)$;
The auxiliary segmentation result of characters with labels $R' = (t'_1, t'_2, t'_3 \cdots t'_n)$;
$\exists R_u \subset R, \exists R_u \subset R'$. Choose all continuous splits that meet the conditions (t_i is the tag (S) and t'_i is not the tag (S and E)), are called rejudged unit R_u;
Threshold value λ;

Output:

Final segmentation result R_o;

$Number = 0$
for t_i in R_u **do**
  Take the maximum co-occurrence frequency f_i of front and back characters.
  $p'_i = \log_{10}(f_i) + \sum_j (p_{ij} * \log p_{ij}) + p_i$
  if $p'_i < \lambda$ **then**
    Remove R^i_u from R_u
  end if
  if $f_i == f_{i-1}$ **then**
    $N_i = Number$
  else
    $N_i = Number + 1$
    $Number = Number + 1$
  end if
end for
Update the R^i_u which has the same N_i

embeddings. Then the embeddings are encoded by the pre-training BERT encoder. The pre-training model is transferred into CWS task through a linear transfer layer. In the end, the marginal probability P_i of each character is computed through a $Softmax$ layer. We only extract the marginal probabilities of characters instead of obtaining the final segmentation result.

2.3 Context-aware Algorithm

The pre-segmenter is more sensitive to word-level splits which are in the system lattice. It does not conjecture the uncertain boundary and leads to a weakness in out-of-vocabulary words. The OOV words are segmented into continuous single words incorrectly. For instance, the word "核苷酸(nucleotide)" is not in the lattice and is labeled as "SSS", which represents three continuous single words. While the auxiliary segmenter that uses character based method is good at conjecturing the dependence between two close characters. The characters "核(core)", "苷(glycosides)" and "酸(acid)" may be predicted as a non-single label by the auxiliary segmenter. Because the "核(core)" and "酸(acid)" have wide ranges of probability as the boundary of the word in this structure, such as "核糖(ribose)" and "核酸(nucleic acid)". Inspired by this idea, some local paths of the pre-segmentation result R are been concerned by the rejudged module. And we utilize a context-aware algorithm to determine the edges of these paths. The algorithm is a global optimization rather than the way to integrate the document-level information into sentence-level optimization. The algorithm is shown as Alg. 1 and R_o represents the final segmentation result.

3 Datasets and Experiments

3.1 Datasets and Settings

To evaluate the performance for CWS on discourse, we propose a new benchmark consisting of two domains (named Chinese daily news and Chinese medical article respectively), as shown in Table 1. A segmentation criterion with fine-grained is adopted, which is close to Peking University (PKU) criterion. This fine-grained criterion is effective in machine translation and prepositional phrase recognition. The training data comes from People's Daily in Jan.1998. The time is far away from the two test data. The size of the new benchmark is shown in Table 1. In this paper, we adopt precision (P), recall (R), and F value to evaluate each method. In addition, the recalls of in-vocabulary (R_{iv}) and out-of-vocabulary (R_{oov}) are considered for evaluation. We choose the median of the range of marginal probability ($p_i \in [0, 1]$) as the threshold $\lambda = 0.5$. The hyper-parameter values in our proposed method is empirical from previous related work (Ma et al., 2018).

3.2 Main Result

We make comprehensive experiments on the new benchmark. We compare the context-aware segmentation with multiple previous proposed methods, which are:

- **Jieba**: a famous Chinese word segmentation tool with domain-specific dictionary. We integrate a medical domain lattice into the tool.
- **LSTM**: an effective and concise model used in Ma et al. (2018). In order to improve the performance on OOV words, we also integrate pre-training embedding into base model.
- **BERT**: a pre-training model with fine-tuning similarly used in Cui et al. (2019).
- **FA-CWS**: a fast and accurate neural method with greedy searching by Cai et al. (2017). The pre-training embedding is based on word2vec.
- **Lattice-LSTM**: a lattice based LSTM with subword encoding proposed by Yang et al. (2019).

In addition, to verify the effectiveness of document-level optimization, we compare our proposed method with the pre-segmenter and the sentence-level method. The sentence-level method does not utilize any document-level information, and the input is a sentence instead of the discourse. The results of Chinese Daily News and Chinese medical articles are shown in Table 2.

From Table 2, it is observed that our proposed method can improve the performance via document-level optimization for Chinese discourse, compared with the methods using the character-level and sentence-level optimization. In addition, the context-aware method does not rely on any external resources. It only extracts the document-level information itself and is not domain limited which is different from previous document segmentation researches. The method is practical for

Corpora		Word	Character	OOV Rate
Train	People's Daily in Jan. 1998	1.2M	2.0M	
Test	Chinese Daily News	42K	63K	6.8%
	Chinese medical articles	40K	66K	27.7%

Table 1: The sizes of the new benchmark

Method	Chinese Daily News					Chinese Medical articles				
	P	R	F	R_{iv}	R_{oov}	P	R	F	R_{iv}	R_{oov}
Jieba	87.76	79.85	83.62	80.15	75.85	85.42	79.60	82.41	78.63	82.13
LSTM	93.12	91.68	92.40	91.80	91.04	86.29	88.49	87.38	90.88	81.66
BERT	93.55	92.49	93.02	93.05	89.17	86.32	87.69	87.00	90.60	79.36
FA-CWS	91.53	90.83	91.18	93.10	60.24	86.94	85.69	86.31	90.13	60.24
Lattice-LSTM	92.86	91.61	92.23	93.13	71.28	79.92	83.70	81.77	89.62	60.58
pre-segmenter	90.53	94.77	92.60	**97.70**	54.69	80.34	89.42	84.64	97.20	67.17
sentence-level	95.22	96.14	95.68	97.45	88.52	83.82	90.84	87.19	**97.63**	71.45
ours	**96.14**	**96.15**	**96.15**	97.02	**91.12**	**87.30**	**92.14**	**89.66**	97.16	**77.77**

Table 2: The result of the new benchmark. The highest values are bold.

the common domain, and it has strong robustness when dealing with a real-world scenario. Compared with previous state-of-the-art character-level and sentence-level works, an obvious improvement is achieved by our proposed method.

Furthermore, due to the ability of external resources, the R_{oov} values of "LSTM" and "BERT" are high when adopting the pre-training embedding. "Jieba" utilizes a medical domain dictionary, and achieves a competitive performance for medical domain segmentation. Our proposed method leverages the information of the test discourse itself to achieve comparable performance.

3.3 Case Study

Existing methods have a potential weakness in dealing with the OOV issue for the new benchmarks. The context-aware method can alleviate this issue for Chinese discourse. Actually, the factors that directly affect the performance of downstream NLP tasks are keywords in discourse. These words have a high probability of OOV words and frequently occur in a document. For instance, the word "萝莉(Lolita)" occurs more than 10 times in the news about Japanese culture. This word is hard to segment in using pre-segmenter because it is not in the lattice. It is segmented into two single words. The auxiliary segmenter pays attention to recognizing this continuous split. Then the context-aware algorithm recalls the splits as one word.

However, it is inevitable that some in-vocabulary words will be affected in the context-aware processing. The value R_{iv} may drop a little in Table 2. For instance, the in-vocabulary word "高山(high mountain)" and "高(Gao)/ 山(Shan)" occurs in discourse at the same time. The two splits represent a common noun and a Chinese person name respectively. Both of them are in the lattice. The Chinese name "高(Gao)/ 山(Shan)" may be segmented as one word incorrectly at the context-aware step. To alleviate this issue, a feasible way is to integrate syntactic knowledge into the model. We will research this idea in the future.

4 Conclusion

In this paper, we intuitively propose a context-aware method to boost the segmentation inconsistency in discourse. The time-efficient and domain knowledge are considered via the document-level information. The method is explainable and unsupervised. In summary, our proposed method is empirical but do not used stiff constraint rules. Besides, a new benchmark in discourse is built for evaluation of the document-level Chinese word segmentation. The distribution of words is natural on benchmark. However, the scale of the benchmark is still limited. We will expand them and make it more reliable. And we will try to integrate the knowledge into popular neural models in the future.

Acknowledgments

We would like to thank the reviewers for their helpful comments and suggestions to improve the qual-

ity of the paper. The authors gratefully acknowledge the financial support provided by the National Key Research and Development Program of China (2020AAA0108004) and the National Natural Science Foundation of China under (No.U1936109, 61672127).

References

Deng Cai, Hai Zhao, Zhisong Zhang, Yuan Xin, Yongjian Wu, and Feiyue Huang. 2017. Fast and accurate neural word segmentation for Chinese. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 608–615, Vancouver, Canada. Association for Computational Linguistics.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Ziqing Yang, Shijin Wang, and Guoping Hu. 2019. Pre-training with whole word masking for chinese bert. *arXiv preprint arXiv:1906.08101*.

Weipeng Huang, Xingyi Cheng, Kunlong Chen, Taifeng Wang, and Wei Chu. 2019. Toward fast and accurate neural chinese word segmentation with multi-criteria learning. *arXiv preprint arXiv:1903.04190*.

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying conditional random fields to japanese morphologiaical analysis. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 230–237, Barcelona, Spain.

Si Li and Nianwen Xue. 2014. Effective document-level features for chinese patent word segmentation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 199–205.

Xiaoya Li, Yuxian Meng, Xiaofei Sun, Qinghong Han, Arianna Yuan, and Jiwei Li. 2019. Is word segmentation necessary for deep learning of chinese representations? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3242–3252.

Ji Ma, Kuzman Ganchev, and David Weiss. 2018. State-of-the-art chinese word segmentation with bi-lstms. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4902–4908.

Yuxian Meng, Wei Wu, Fei Wang, Xiaoya Li, Ping Nie, Fan Yin, Muyu Li, Qinghong Han, Xiaofei Sun, and Jiwei Li. 2019. Glyce: Glyph-vectors for chinese character representations. In *Advances in Neural Information Processing Systems*, pages 2742–2753.

Yuanhe Tian, Yan Song, Xiang Ao, Fei Xia, Xiaojun Quan, Tong Zhang, and Yonggang Wang. 2020. 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*, pages 8286–8296.

Nianwen Xue. 2003. Chinese word segmentation as character tagging. In *International Journal of Computational Linguistics & Chinese Language Processing, Volume 8, Number 1, February 2003: Special Issue on Word Formation and Chinese Language Processing*, pages 29–48.

Qian Yan, Chenlin Shen, Shoushan Li, Fen Xia, and Zekai Du. 2017. Domain-specific chinese word segmentation with document-level optimization. In *National CCF Conference on Natural Language Processing and Chinese Computing*, pages 353–365. Springer.

Jie Yang, Yue Zhang, and Fei Dong. 2017. Neural word segmentation with rich pretraining. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 839–849, "Vancouver, Canada". "Association for Computational Linguistics".

Jie Yang, Yue Zhang, and Shuailong Liang. 2019. Subword encoding in lattice lstm for chinese word segmentation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2720–2725.

Yuxiao Ye, Weigang Li, Yue Zhang, Likun Qiu, and Jian Sun. 2019. Improving cross-domain chinese word segmentation with word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2726–2735.

Qi Zhang, Xiaoyu Liu, and Jinlan Fu. 2018. Neural networks incorporating dictionaries for chinese word segmentation. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Hai Zhao, Chang-Ning Huang, Mu Li, and Bao-Liang Lu. 2010. A unified character-based tagging framework for chinese word segmentation. *ACM Transactions on Asian Language Information Processing*, 9(2):1–32.

Hai Zhao and Chunyu Kit. 2008. Unsupervised segmentation helps supervised learning of character tagging for word segmentation and named entity recognition. In *The Sixth SIGHAN Workshop on Chinese Language Processing*, pages 106–111, Hyderabad, India.

Lujun Zhao, Qi Zhang, Peng Wang, and Xiaoyu Liu. 2018. Neural networks incorporating unlabeled and partially-labeled data for cross-domain chinese word segmentation. In *IJCAI*, pages 4602–4608.

Long Zhou, Wenpeng Hu, Jiajun Zhang, and Chengqing Zong. 2017. Neural system combination for machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 378–384, Vancouver, Canada. Association for Computational Linguistics.