Extra Large Sequence Transformer Model for Chinese Word Segment

Dezhou Shen
Department of Computer Science
Tsinghua University
Beijing, CN 100084
sdz15@mails.tsinghua.edu.cn

Abstract

Chinese word segment is widely studied in document analysis. The accuracy of the current popular word segment model, LSTM+CRF, is still not satisfactory. Models trained by the popular dataset often fails in the out-domain situation. In this paper, combining the Transformer-XL layer, the Fully-Connect layer, and the Conditional Random Field layer, the proposed model improved 3.23% in the macro-F1 score, comparing to the BERT+CRF model, on the MSR2005 Chinese word segment test dataset.

1. Introduction

As a fundamental study for the semantic document segment, the Chinese word segment has been studied for decades.

By tagging the status of each token, intended entities with longer than one character could be drawn out of documents, and unintended tokens are assigned with a special token type which are normally one character length.

2. Related Work

Devlin et al. [1] proposed Bidirectional Encoder Representations from Transformers(BERT), and achieved 92.8% on the CoNLL-2003 test set by fine-tuning BERT on the downstream Named Entity Recognition task.

Yang et al. [2] proposed an autoregressive language model which overcomes the limitations of BERT in sequence length, results showed that XLNet achieved promising results on question answering, natural language inference, sentiment analysis, and document ranking tasks.

Cui et al. [3] fine-tuned BERT and XLNet with extra 4.5 billion Chinese tokens, and improved the performance of many downstream tasks.

Tian et al. [4] showed that the extra pre-trained BERT language model as encoder concatenating to the Conditional Random Fields(CRF) can achieve 98.28% to 98.40% on the MSR2005 Chinese word segment test dataset.

Cui and Zhang [5] proposed a hierarchically-refined label attention network in Named Entity Recognition task, the result showed that LSTM+CRF is less competitive than their LSTM+LAN model on the WSJ, Universal Dependencies v2.2, OntoNotes 5.0, and CGGBank dataset.

Lafferty et al. [6] proposed the CRF probabilistic model to segment and label sequence data, and outperforms the hidden Markov models and maximum entropy Markov models by 0.14% 0.82% in accuracy on the Penn treebank POS dataset.

3. Methodology

3.1. Data Augment

For the Chinese 8-bit Unicode Transformation Format encodings, there are two different encoding for the same printable ASCII tokens, one is named as the half character encoding, and the other is named as the full character encoding. Their difference relies on that one token take one byte or three bytes. The MSR Chinese word segment dataset[9] used the full character encoding, but in most cases, people tend to use the half character encoding to comply with ASCII code for compatibility reason. In this paper, I use a conversion function to convert the MSR2005 training dataset tokens to half character encoding for the tokens which have the corresponding tokens in the ASCII table. Thanks to the data augment method, training dataset sentences size is increased from 86,924 to 166,740, and moreover, models trained shall perform well in both the half character and full character corpus. Data augment solves the problem that when the dataset are full character encoding characters, however, most use cases are mixed character encodings by half and full.

3.2. Tokenization, Encoding And Labeling

For a document with multiple sentences and paragraphs, the first step is splitting the document into lines by carefully choosing the separator token. For Chinese sentences, the line-breaks, commas, periods, question marks, and exclama-
tion marks are the most common separator tokens. Before encoding the tokens, sentences are padded with padding token [0] to the max sequence length. I choose 128 as the max sequence length for the MSR2005 dataset. After tokenization, I use the XLNet encoder to represent each token as the token integers. Each token has been labeled according to its type, for the one character entity, its status is labeled to [0]. For the longer than one character entity, the first character is labeled to [1], the last character is labeled to [3], and the rest characters are labeled to [2].

3.3. Sequence Encoder

After the encoding steps, the tokens are encoded to token integers which represents their vocabulary indices. Transformer-XL[7] also named as XLNet, a novel neural architecture that enables learning dependency beyond a fixed length without disrupting temporal coherence. To encode by XLNet and BERT model, the mask tokens and the segment tokens are prepared. The non-padding tokens’ mask tokens are set to [1], and the segment tokens are set to [0]. The padding tokens’ make tokens are set to [0], and the segment tokens are set to [1] for telling the model the boundary of the sentences. For encoding the sequence, all the layers are preserved for later usage. The vocabulary encoding used by the LSTM word embedding doesn’t need the mask tokens and segment tokens, the vocabulary was composites by characters extracted from the MSR2005 training dataset, its conversion from character set to half character encoding set, and one padding token [PAD], and one unknown token [UNK]. There are 5245 tokens in the vocabulary, and the embedding size is 256.

3.4. Sequence Hidden Representation

For a naive solution, it is perfect to add a softmax layer, or a CRF layer. However, I designed multiple layers to take advantage of the different layers for the transfer learning. Firstly, take the first layer, and the last layer of the XLNet sequence encoding. Then use two dropout layers with each dropout rate of 5%, which means the probability of an element to be zeroed. Thirdly, I use a fully connected layer to map the hidden size of XLNet to the status types, which is 4 in the Chinese word segment task. Lastly, take the logarithm, and a softmax conversion before concatenating to the CRF layer.

3.5. Conditional Random Field

Conditional Random Field is a traditional method to keep the output of the sequence following a reasonable pattern. Each token sequence has a corresponding mask sequence to show its margin, the token [True] shows that the token is a non-padding token, and the token [False] shows that the token is a padding token.

3.6. Additional Loss function

Besides, the Conditional Random Field loss for training, I added a type loss for training speed. After the two dropout layers, an extra fully connected layer was concatenated to the last dropout layer to get the type loss. Then use a multi-label one-versus-all loss based on max-entropy to get the sequence type loss. The type sequence target was smoothed by 0 and 1 values, the type [0] for the token status of [0], and the type [1] for the other token status.

3.7. Decoder

For the decoding steps, if the status is [0], then the entity is the corresponding token. If the status is [1], it shows that the token belongs to a new entity, and if the status is [2], it tells that the token belongs to current working entity, and if the status is [3], it tells that the token belongs to current working entity and the entity reached an end token. When decoding, I use a CRF decoder, which benefits from the state transfer matrix and avoids error state sequence during the decoding steps. The decoder used the Viterbi algorithm[8] which make the decoding steps faster.

3.8. Summary

As Figure (1) shows, the Chinese word segment architecture are formed by input layer, encoder layer, concatenating layer, representation layer, CRF layer, and output layer. In this paper, I mainly focus on discussing the improvement on the encoder layer.

4. Experiments

4.1. Dataset

The MSR Chinese word dataset has 2.37 million words in the training set with 88 thousand vocabulary size, and has 107 thousand words in the test set with 13 thousand vocabulary size. As Emerson[9] reported that, during the second international Chinese word segment competition, the best F1-score achieved was 97.2% with permission of using the additional corpus on the MSR test dataset.

4.2. Common Setups

The experiment was performed on a Google Cloud NVIDIA V100 with 16GB of the high speed of memory each, with 32GB of RAM, which can provide 112 tera-flops of computation capability.

The learning rate is $4^{-5}$, and the batch size is 64, for the BERT and XLNet. The learning rate is $2^{-3}$, and the batch size is 512 for the bidirectional LSTM. The max sequence length is 128, and the training epoch is 10. For the LSTM model, I use the word embedding layer with embedding size of 256, and the bidirectional LSTM layers with 384 neurons in each layer. I use the popular pre-trained RoBERTa
language models trained on 4.5 billion of Chinese[3], and use the pre-trained XLNet language models trained on 4.5 billion of Chinese tokens.

All the Table (1) models share almost the same network architecture except for the embedding of the tokens.

4.3. Reproducibility

The data augment, the LSTM, RoBERTa, XLNet, CRF codes and pre-trained Chinese word segment models are available at https://drive.google.com/drive/folders/1-GwAkpE7qQs0Im0PqD6a26T0vVrxJg3mR. And the supplementary codes are available at https://github.com/ShenDezhou/EXLNet.

5. Conclusion

In this paper, I introduced a novel Chinese word segment architecture, which leveraging hierarchical layers of the document encoding layers. By keeping the Conditional Random Field layer, I proposed a novel sequence hidden representation of the document sequence, which achieved 3.23% performance gain in macro F1-score than the BERT and CRF architecture on the MSR2005 Chinese word segment test dataset. Not surprisingly, the performance of the proposed XLNet-CRF model also exceed the best result of the Second International Chinese Word Segmentation Competition.

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, & Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, 2019.

[2] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, & Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. In Advances in neural information processing systems, pages 5753–5763, 2019.

[3] Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, & Guoping Hu. Revisiting pre-trained models for Chinese natural language processing. In Findings of EMNLP. Association for Computational Linguistics, 2020.

[4] Yuanhe Tian, Yan Song, Fei Xia, Tong Zhang, & Yonggang Wang. Improving Chinese word segmentation with wordhood memory networks. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8274–8285, 2020.
[5] Leyang Cui & Yue Zhang. Hierarchically-refined label attention network for sequence labeling. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 4106–4119, 2019.

[6] John D. Lafferty, Andrew McCallum, & Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings of the Eighteenth International Conference on Machine Learning, ICML ’01, page 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[7] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G Carbonell, Quoc Le, & Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 2978–2988, 2019.

[8] G David Forney. The viterbi algorithm. Proceedings of the IEEE, 61(3):268–278, 1973.

[9] Thomas Emerson. The second international chinese word segmentation bakeoff. In Proceedings of the fourth SIGHAN workshop on Chinese language Processing, 2005.