Unsupervised Word Segmentation with BERT Oriented Probing and Transformation

Wei Li[1]*, **Yuhan Song**[2]*, **Qi Su**[3], **Yanqiu Shao**[1]
[1]School of Information Science, Beijing Language and Culture University
[2]School of EECS, Peking University
[3]School of Foreign Languages, Peking University
liweitj47@blcu.edu.cn
{songyuhan, sukia}@pku.edu.cn
shaoyanqiu@blcu.edu.cn

Abstract

Word Segmentation is a fundamental step for understanding many languages. Previous neural approaches for unsupervised Chinese Word Segmentation (CWS) only exploit shallow semantic information, which can miss important context. Large scale Pre-trained language models (PLM) have achieved great success in many areas. In this paper, we propose to take advantage of the deep semantic information embedded in PLM (e.g., BERT) with a self-training manner, which iteratively probes and transforms the semantic information in PLM into explicit word segmentation ability. Extensive experiment results show that our proposed approach achieves a state-of-the-art F1 score on two CWS benchmark datasets. The proposed method can also help understand low resource languages and protect language diversity.[1]

1 Introduction

There exist many low resource fields and languages where labeled word segmentation is inaccessible, which makes unsupervised word segmentation desirable. Previous unsupervised word segmentation methods mainly apply statistical models to either evaluate the quality of possible segmented sequence with discriminative models (e.g., Mutual Information (Chang and Lin, 2003)) or estimate the generative probabilities with generative models (e.g., Hidden Markov Model (Chen et al., 2014)). However, these statistical methods can only make use of the limited contextual information, thus yielding less competitive performance.

With the thrive of neural networks, researchers have applied neural models for unsupervised word segmentation. Sun and Deng (2018) propose a segmental language model (SLM) to estimate the generative probability with recurrent networks. Although SLM can exploit more contextual information compared with statistical models, it is still weak in modeling deep semantic information, limited by its model capacity and training data scale.

Pre-trained language models trained on large scale data have shown superior ability to model contextual information, and achieve great success in various tasks (Peters et al., 2018; Devlin et al., 2019; Radford et al., 2019). Inspired by the attempt for interpreting BERT (Wu et al., 2020), we propose to take advantage of the semantic representation ability of BERT to evaluate the closeness between characters in a **probing manner**. To be more specific, we assume that the difference between masking one character and masking several adjacent characters as a whole reveals the closeness between that character and the adjacent ones.

Although this probing-based method can take advantage of the large amount of knowledge embedded in BERT, it only implicitly exploits the representation ability of BERT. To transfer the implicit knowledge into explicit segmentation boundary, we propose to apply a self-training method that **transforms** the segmentation decision from generative methods with high confidence into traditional "BI" sequence labeling system, which is then treated as the supervision signals for a discriminative model.

To combine the advantage of both generative and discriminative models, we propose to iteratively train the discriminative model and generative model under the supervision signal from their counterparts. To select the model with the best performance in the unsupervised setting, we propose an evaluation module that evaluates the quality of the word boundaries with masked prediction accuracy based on the assumption that the closer two characters are, the bigger loss masking one adjacent character would bring.

We conduct experiments on two Chinese Word Segmentation benchmark datasets in an unsuper-

*Equal Contribution.

[1] The code is available at https://github.com/liweitj47/BERT_unsupervised_word_segmentation

Findings of the Association for Computational Linguistics: ACL 2022, pages 3935 - 3940
May 22-27, 2022 ©2022 Association for Computational Linguistics

vised manner. Experiment results show that our method can outperform the strong baseline models and achieve state-of-the-art results in unsupervised CWS. Extensive analysis shows the effectiveness of the proposed modules.

We conclude our contributions as follows:

- We propose an unsupervised word segmentation method that segments tokens by probing and transforming PLM with generative and discriminative modules, which are trained in a mutual promotion manner and selected for inference with an evaluation module.

- Experiment results show that our proposed method achieves the state-of-the-art result in unsupervised CWS. Extensive analysis testifies the effectiveness of the proposed modules.

2 Related Work

Previous unsupervised word segmentation methods can be roughly classified as generative and discriminative two ways. Generative models focus on finding the segmented sequence with the highest posterior probability. Hierarchical Dirichlet process (HDP) model (Goldwater et al., 2009), Nested PitmanYor process (NPY) (Mochihashi et al., 2009), Hidden Markov Model (HMM) (Chen et al., 2014) and SLM (Sun and Deng, 2018) are all different ways to estimate the generative probabilities for segmented sequences. On the other hand, discriminative models focus on designing a measure to evaluate the segmented sequences. Mutual Information (MI) (Chang and Lin, 2003), normalized Variation of Branching Entropy (nVBE) (Magistry and Sagot, 2012) and ESA (Wang et al., 2011) apply co-occurrence based measurement to evaluate the segmented sequences.

3 Approach

In this section, we describe our BERT oriented probing and transformation based unsupervised word segmentation approach. Our model mainly consists of three parts, a **generative** module that suggests the plausible word boundaries by **probing** BERT, a **discriminative** module that **transforms** the implicit boundary information into explicit sequence labels, and an **evaluation** module that estimates the performance of the model in an unsupervised manner.

Algorithm 1 Unsupervised Word Segmentation Procedure

```
Require: Generative Module $G$, Discriminative Module $D$, Evaluation Module $E$, sequences to be segmented $X$.
  i = 0
  while True do
    Segment the sequences $X$ with $G$ into $X^g$
    Transform the segmented $X^g$ into “BI” labels
    Train $D$ with high confident segmentations in $X^g$
    Segment the sequences $X$ with updated $D$ into $X^d$
    Train $G$ with high confident segmentations in $X^d$
    Evaluate the segmented sequence $X^d$ with $E$
    $e = E(X^d)$
    if $e^i < e^{i-1}$ then
      Return $D^{i-1}$
    end if
    i += 1
  end while
```

3.1 Overview

Because our method works in an unsupervised manner, we propose to get the original word boundary information by probing BERT, which reveals the word boundaries by measuring the distance between masking a span and masking a token using the generative module. This distance reflects the closeness between the masked token and the masked span (separately). Then the discriminative module transforms the word boundaries suggested by the generative module into explicit segmentation labels to enable the self-training process. To combine the advantages of both generative and discriminative modules, two modules are iteratively trained with the word boundaries suggested by the updated counterpart with high confidence. To decide when to stop this iterative self-training procedure, an evaluation module is proposed to evaluate the segmented sequence, which early stops the iterative process with the model parameters that yields the best performance.

3.2 Generative Module

The proposed generative module works by probing a pre-trained language model (e.g., BERT) with masks on tokens. Assume the input sequence to be $[x_1, x_2, \cdots, x_n]$. We first mask one token at a time in order. The representation at i-th position given by BERT after masking x_i is H_i. Then we mask two successive tokens at a time in order. $H_{i,j}$ is the representation given by BERT at i-th position after masking both x_i and x_j. Note that it is different for the representation at j-th position after masking both x_i and x_j, which we denote as $H_{j,i}$.

The intuition behind the generative model is that we assume if two tokens x_i, x_j are inherently close and should be combined as a word, the difference between masking i-th and j-th token together and solely masking i-th token should be large, which is reflected by the probing distance d,

$$d = \frac{(|H_{i,j} - H_i| + |H_{j,i} - H_j|)}{2}$$

On the contrary, if two tokens are loosely connected, d should be small. This assumption follows the intuition that if x_i is largely dependent on x_j, masking x_j should bring a relatively big influence on the representation.

This indicator is applied to segment token sequence with a threshold, that is to say, if $d \geq threshold$, we combine the two tokens x_i and x_j, if $d \leq threshold$, we segment x_i and x_j.

3.3 Discriminative Module

The generative module can only exploit the implicit segmentation revealed by BERT. Furthermore, it is not very friendly when the word length gets longer. To overcome these drawbacks, we propose to transform the segmentation information provided by the generative module with high confidence into traditional supervised sequence labeling scheme with "BI" labels, which indicates the role (position) of the token to be "beginning" ("B") or "inside" ("I") of a word.

We train the discriminative module by fine-tuning BERT on the transformed sequence labels with an additional output layer projecting the representation into "BI" labels. Since the results given by the generative module can be noisy, we only adopt the combined words with relatively high confidence, which is realized by strict thresholds for the generative module. If $d \geq threshold_l$, we combine the two tokens x_i and x_j, if $d \leq threshold_h$, we segment x_i and x_j. $threshold_l$ indicates lower bound, $threshold_h$ indicates higher bound.

3.4 Iterative Training and Evaluation Module

We assume that the generative module and the discriminative module can capture segmentation information from different aspects. Therefore, we propose a self-training procedure, which promotes both the generative module and the discriminative module by making them learn from the high confident predictions of the counterpart.

To make the generative module learn from the discriminative module, we design a Euclidean distance based MSE loss function

$$loss_{generative} = \|d - threshold\|^2$$

to push the distance between two tokens predicted to be in the same word to be larger than a threshold and vice versa. The loss is activated only when the generative module makes different predictions from the discriminative module.

To prevent the self-training procedure from being over-fitting, we propose to keep the MLM objective while fine-tuning on the word segmentation objectives, and early stop the training with an evaluation module. The intuition behind the evaluation module is that predicting a masked token with the token inside the same word is much easier than predicting this masked token with the token outside that word. Formally, let the cross-entropy of predicting the i-th token x_i with the masked language modeling ability of BERT when masking two adjacent tokens $x_{i,j}$ be $CE_{i,j}$, we assume that

$$CE_{i-1,i} < CE_{i,i+1}$$

if $x_{i,i+1}$ rather than $x_{i-1,i}$ belongs to the same word, because x_{i+1} provides more information for prediction when masking $x_{i-1,i}$.

We apply this principle to inspect the segmentation results from either the discriminative module or the generative module. When the evaluation module detects performance decline, the training procedure stops, and the discriminative module with the best performance is used as the final word segmentation model.

4 Experiment

In this section, we show the results and analysis on two CWS benchmark datasets, PKU and MSR for a fair comparison, which are provided by the Second Segmentation Bake-off (SIGHAN 2005) (Emerson, 2005). There are 104K and 107K words in the test set of PKU and MSR datasets respectively.

4.1 Settings

In this paper, we use the pre-trained BERT (base) model for Chinese and the corresponding tokenizer

F1 score	PKU	MSR
HDP (Goldwater et al., 2009)	68.7	69.9
NPY-3 (Mochihashi et al., 2009)	-	80.7
NPY-2 (Mochihashi et al., 2009)	-	80.2
ESA (Wang et al., 2011)	77.8	80.1
nVBE (Magistry and Sagot, 2012)	80.0	81.3
HDP + HMM (Chen et al., 2014)	75.3	76.3
Joint (Chen et al., 2014)	81.1	81.7
SLM-2 (Sun and Deng, 2018)	80.2	78.5
SLM-3 (Sun and Deng, 2018)	79.8	79.4
MSLM (Downey et al., 2021)	62.9	-
Proposal	**84.1**	**83.0**

Table 1: F1 score on two word segmentation benchmark datasets. Our proposed method achieves the state-of-the-art performance on all the datasets. We take the results reported in the original paper.

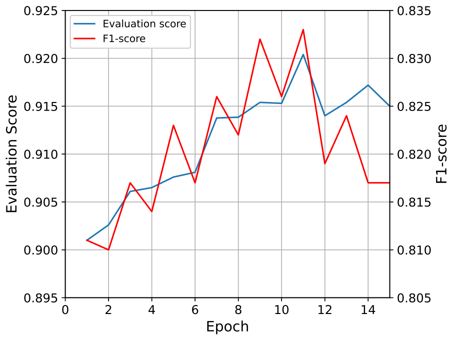

Figure 1: The relation between evaluation score and F1 score on the development set. The evaluation score shows good coherence with F1 score. We select the model with best evaluation score, which also achieves the best F1 score on the development set.

released by Huggingface.[2] The tokenizer tokenizes the sentence into Chinese characters, which involves with no word (segmentation) information. We randomly initialize the discriminative module, which is trained for 2 epochs using sequence labels transformed from the generative module with high confidence. $threshold_l$ is 8 and $threshold_h$ is 12. We use AdamW (Loshchilov and Hutter, 2019) optimizer with the learning rate of 1e-4.

4.2 Results

In Table 1 we show the F1 score on PKU and MSR. From the results, we can see that our model yields much better results than the previous models and achieves state-of-the-art results. We assume the reason behind is that our model can take advantage of the large pre-trained language model, which encodes abundant language matching knowledge and can better model the context with big model capacity. Moreover, we can observe that the neural-based model *SLM* does not outperform the traditional statistical *Joint* method, but gives better results than other traditional generative models. This indicates that combining generative and discriminative methods can benefit the results. Moreover, our model does not need to constrain the longest word length compared with SLM-2, SLM-3, etc., which provides more flexibility. This is achieved by introducing the discriminative module, which segments the words under the sequence labeling scheme.

F1 score	PKU	MSR
Generative Only	74.8	72.5
+Discriminative	79.7	78.3
+Discriminative & iterative	80.5	78.9
+Discriminative & mlm	82.0	82.1
Full Model	84.1	83.0

Table 2: Ablation study results. "mlm" means using mlm loss as a regularization mentioned in Section 3.4. "iterative" means using iterative training mentioned in Section 3.4. "Full model" means using Discriminative & mlm & iterative training.

4.3 Ablation Study

In Table 2 we show the results for removing the designed modules. "Generative only" means we only use the generative module described in section 3.2, where a hard threshold of 10 is used to decide the word boundary. "+Discriminative" means we use the discriminative module after learning from the generative module described in section 3.3 without iterative training and mlm loss. From the results, we can see that revealing the implicit word boundary information by probing BERT can only provide performance comparable to traditional statistical models. Transforming the implicit knowledge into explicit segmentation labels (+Discriminative) can give big promotion, which makes better use of the big amount of semantic knowledge encoded in PLM. Moreover, the proposed iterative training process and mlm loss further help improve the overall performance by combining the advantages of both generative and discriminative modules.

Effect of Evaluation Module In Figure 1, we show the relation between the evaluation score described in section 3.4 and the development F1 score. We can see that the model with the best evaluation score achieves the best F1 score in the development set, and it generally coordinates with the variation trend of the F1 score, which makes the evaluation

[2] https://huggingface.co/bert-base-chinese

score a reasonable indicator to select the best model in the unsupervised setting.

4.4 Case Study

In Table 3 we show a concrete example of the segmentation results of SLM and our proposed method. Both two methods basically give correct word segments. The disagreement mainly lies in “送交市政府” (give to the city government). Compared with other words, “ 送交” can be relatively rare and bears very similar meaning with the single character “送”, which makes SLM wrongly segment “送交” apart. On the contrary, our method is built based on BERT trained on a large corpus, which makes our model able to recognize these relatively rare words. For the part “市政府”, where our model chooses to split, we assume that this is because similar contexts are often seen such as “ 北京市” (Beijing City), where “市” should be separated from “政府” (government). Furthermore, separating “市政府” into two words does not affect the understanding of the original text, and is more dependent on the segmentation fineness.

5 Conclusion

In this paper, we propose a BERT oriented Probing and Transformation method for unsupervised Word Segmentation. Our proposed model reveals the semantic information encoded in PLM into word boundary information by probing and transforming the token representations into explicit sequence labels. Experiment results on two benchmark CWS datasets show that our method achieves state-of-the-art F1 score. The proposed method works in an unsupervised manner, which can help understand low resource and endangered languages and thus protecting language diversity.

Acknowledgements

This research project is supported by National Key R&D Program of China (No. 2019YFC1521200), the National Natural Science Foundation of China (No. 61872402), the Humanities and Social Science Project of the Ministry of Education (No. 17YJAZH068), Science Foundation of Beijing Language and Culture University (supported by “the Fundamental Research Funds for the Central Universities”) (No. 21YBB19)

References

Jason S. Chang and Tracy Lin. 2003. Unsupervised word segmentation without dictionary. In *ROCLING 2003 Poster Papers*, pages 355–359, Hsinchu, Taiwan. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

Miaohong Chen, Baobao Chang, and Wenzhe Pei. 2014. A joint model for unsupervised Chinese word segmentation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 854–863, Doha, Qatar. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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, Minneapolis, Minnesota. Association for Computational Linguistics.

C. Downey, Fei Xia, Gina-Anne Levow, and Shane Steinert-Threlkeld. 2021. A masked segmental language model for unsupervised natural language segmentation. *ArXiv*, abs/2104.07829.

Thomas Emerson. 2005. The second international Chinese word segmentation bakeoff. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*.

Sharon Goldwater, Thomas L. Griffiths, and Mark Johnson. 2009. A bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, 112(1):21–54.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Pierre Magistry and Benoît Sagot. 2012. Unsupervized word segmentation: the case for Mandarin Chinese. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 383–387, Jeju Island, Korea. Association for Computational Linguistics.

Daichi Mochihashi, Takeshi Yamada, and Naonori Ueda. 2009. Bayesian unsupervised word segmentation with nested Pitman-Yor language modeling. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 100–108, Suntec, Singapore. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of*

Model	Segmentation
Gold	她 保证 ， 学生 们 的 意见 将 送交 市政府 领导 机关 。
SLM	她 保证 ， 学生 们 的 意见 将 送 交市 政府 领导 机关 。
Proposal	她 保证 ， 学生 们 的 意见 将 送交 市 政府 领导 机关 。

Table 3: Segmentation results of SLM and our proposed method. The gold content can be loosely translated as “She proposed that the suggestions of the students would be transferred to the leading agency of the city government.”

the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Zhiqing Sun and Zhi-Hong Deng. 2018. Unsupervised neural word segmentation for Chinese via segmental language modeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4915–4920, Brussels, Belgium. Association for Computational Linguistics.

Hanshi Wang, Jian Zhu, Shiping Tang, and Xiaozhong Fan. 2011. A new unsupervised approach to word segmentation. *Computational Linguistics*, 37(3):421–454.

Zhiyong Wu, Yun Chen, Ben Kao, and Qun Liu. 2020. Perturbed masking: Parameter-free probing for analyzing and interpreting BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4166–4176, Online. Association for Computational Linguistics.