Unsupervised Word Segmentation with BERT Oriented Probing and Transformation

Anonymous ACL submission

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 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 unsupervised 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.
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 $e < e^{i-1}$ then
    Return $D^{i-1}$
  end if
  
  $i += 1$
end while

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 Pitman-Yor 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.

3.1 Overview

Our model starts by probing BERT, which yields the distance between masking a span and masking a token using the generative module. Then the discriminative module transforms the word boundaries suggested by the generative module into explicit segmentation labels. 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 applied to evaluate the segmented sequence, which stops the iterative process and yields the model with 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 $x_j$. We assume that if two tokens $x_i, x_j$ are inherently close and should be combined as a word, the probing distance

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

should be large. On the contrary, if two tokens are loosely connected, $d$ should be small. This assumption follows the intuition that if $x_j$ 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 $d \geq \text{threshold}$, we combine the two tokens $x_i$ and $x_j$, if $d \leq \text{threshold}$, we segment $x_i$ and $x_j$.

3.3 Discriminative Module

Because the generative module can only exploit the implicit segmentation revealed by BERT, we
### 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.

<table>
<thead>
<tr>
<th>Method</th>
<th>F1 score</th>
<th>PKU</th>
<th>MSR</th>
</tr>
</thead>
<tbody>
<tr>
<td>HDP (Goldwater et al., 2009)</td>
<td>68.7</td>
<td>69.9</td>
<td></td>
</tr>
<tr>
<td>NPY-3 (Mochihashi et al., 2009)</td>
<td>-</td>
<td>80.7</td>
<td></td>
</tr>
<tr>
<td>NPY-2 (Mochihashi et al., 2009)</td>
<td>-</td>
<td>80.2</td>
<td></td>
</tr>
<tr>
<td>ESA (Wang et al., 2011)</td>
<td>77.8</td>
<td>80.1</td>
<td></td>
</tr>
<tr>
<td>nVBE (Magistry and Sagot, 2012)</td>
<td>80.0</td>
<td>81.3</td>
<td></td>
</tr>
<tr>
<td>HDP + HMM (Chen et al., 2014)</td>
<td>75.3</td>
<td>76.3</td>
<td></td>
</tr>
<tr>
<td>Joint (Chen et al., 2014)</td>
<td>81.1</td>
<td>81.7</td>
<td></td>
</tr>
<tr>
<td>SLM-2 (Sun and Deng, 2018)</td>
<td>80.2</td>
<td>78.5</td>
<td></td>
</tr>
<tr>
<td>SLM-3 (Sun and Deng, 2018)</td>
<td>79.8</td>
<td>79.4</td>
<td></td>
</tr>
<tr>
<td>MSLM (Downey et al., 2021)</td>
<td>62.9</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td><strong>Proposal</strong></td>
<td><strong>84.1</strong></td>
<td><strong>83.0</strong></td>
<td><strong>5</strong></td>
</tr>
</tbody>
</table>

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.

---

proposing to transform the segmentation information provided by the generative module with high confidence into traditional supervision sequence labels “BI”, which indicates the role of the token to be “beginning” or “inside” of a word. We train the discriminative module by fine-tuning BERT on the transformed labels with an additional output layer projecting the representation into “BI” labels. High confidence is realized by strict thresholds. If $d \geq \text{threshold}_i$, we combine the two tokens $x_i$ and $x_j$, if $d \leq \text{threshold}_h$, we segment $x_i$ and $x_j$. $\text{threshold}_i$ indicates lower bound, $\text{threshold}_h$ indicates higher bound.

### 3.4 Iterative Training and Evaluation Module

We assume that generative module and discriminative module can capture different segmentation knowledge. Therefore, we propose a self-training procedure, which makes the generative module and discriminative module learn from the high confident predictions of each other. To make the generative module learn from the discriminative module, we design a Euclidean distance based loss function

$$\text{loss}_{\text{generative}} = ||d - \text{threshold}||^2$$

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 effective 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 training and 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 $x_{i,i+1}$ provides more information for prediction when masking $x_{i-1,i}$.

We apply this principle to inspect the segmentation result 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 released by Huggingface. We randomly initialize the discriminative module, which is trained for 2 epochs using sequence labels transformed from the generative module with high confidence. $\text{threshold}_i$ is 8 and $\text{threshold}_h$ is 12. We use AdamW (Loshchilov and Hutter, 2019) optimizer with the learning rate of 1e-4.

1<https://huggingface.co/transformers/pretrained_models.html>
Table 2: Segmentation results of SLM and our proposed method.

<table>
<thead>
<tr>
<th>Model</th>
<th>Segmentation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gold</td>
<td>她 保证，学生们的意见将送交市政府领导机关。</td>
</tr>
<tr>
<td>SLM</td>
<td>她 保证，学生们的意见将送交市政府领导机关。</td>
</tr>
<tr>
<td>Proposal</td>
<td>她 保证，学生们的意见将送交市政府领导机关。</td>
</tr>
</tbody>
</table>

Table 3: Ablation study results.

<table>
<thead>
<tr>
<th>F1 score</th>
<th>PKU</th>
<th>MSR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full Model</td>
<td>84.1</td>
<td>83.0</td>
</tr>
<tr>
<td>Generative Only</td>
<td>74.8</td>
<td>72.5</td>
</tr>
<tr>
<td>Generative+Discriminative</td>
<td>82.0</td>
<td>82.1</td>
</tr>
</tbody>
</table>

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.

4.3 Ablation Study

In Table 3 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. “Generative+Discriminative” means we use the discriminative module after learning from the generative module described in section 3.3 without iterative training. 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 knowledge encoded in PLM. Moreover, the proposed iterative training process further helps 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 score a reasonable indicator to select the best model in the unsupervised setting.

4.4 Case Study

In Table 2 we show one 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.
References


