A Prism Module for Semantic Disentanglement in Name Entity Recognition

Kun Liu[1,*,†] **Shen Li**[2,*] **Daqi Zheng**[2] **Zhengdong Lu**[2] **Sheng Gao**[1] **Si Li**[1]

{liukun, gaosheng, lisi}@bupt.edu.cn
{shen, da, luz}@deeplycurious.ai

[1] Beijing University of Posts and Telecommunications
[2] Deeplycurious.ai

Abstract

Natural Language Processing has been perplexed for many years by the problem that multiple semantics are mixed inside a word, even with the help of context. To solve this problem, we propose a prism module to disentangle the semantic aspects of words and reduce noise at the input layer of a model. In the prism module, some words are selectively replaced with task-related semantic aspects, then these denoised word representations can be fed into downstream tasks to make them easier. Besides, we also introduce a structure to train this module jointly with the downstream model without additional data. This module can be easily integrated into the downstream model and significantly improve the performance of baselines on named entity recognition (NER) task. The ablation analysis demonstrates the rationality of the method. As a side effect, the proposed method also provides a way to visualize the contribution of each word. [1]

1 Introduction

In Nature Language Processing (NLP), words contribute differently to different tasks. Therefore, attention-based models pay more attention on important words than unimportant words. Since the information that is unrelated to the task can be regarded as noise, unimportant words contain more noise than important words do. From this perspective, attention is a noise reduction mechanism.

Hard attention and soft attention are two main types of attention mechanisms which are proposed in (Xu et al., 2015). Hard attention mechanism selects some important tokens from input sequence and ignore others. This will lead to the loss of necessary information which exists in the ignored tokens. By contrast, in soft attention mechanism, a probability distribution which reflects the importance of tokens is calculated over each token of the input sequence. However, since there is more useless information than useful information in unimportant words, it should be noted that noise could be kept more, when those words are assigned with non-zero probabilities. Overall, both two attention mechanisms have drawbacks in noise reduction.

Attention mechanism is firstly applied in Computer Vision (CV) (Mnih et al., 2014) where pixels are the basic units. However, in NLP, the minimum unit is not word but sense. Therefore, NLP tasks need a noise reduction method at a finer granularity than attention mechanism.

Normally, various aspects of semantics are entangled in word embeddings (Bengio et al., 2003; Mikolov et al., 2013). However, only some of the aspects are needed in specific tasks and other redundant aspects can be regarded as noise. To reduce the noise, entangled word embeddings can be replaced with distributed representations of disentangled semantic aspects. Considering that it could be hard to find the corresponding semantics for each aspect, we call them abstract aspects.

In this paper, we propose a prism module to generate parallel denoised sentences from multiple aspects. Different from attention mechanism, the module reduces noise in semantic aspect level rather than word level. Specifically, we selectively replace some words in the sentence with abstract aspects. These denoised sentences are expected to keep sufficient information to make predictions in the downstream tasks, like the low-noise version of original sentence. Compared with attention mechanism, the proposed method not only reduces the noise, but also reduces the loss of necessary information. Furthermore, this method also allows

[*] Kun Liu and Shen Li contributed equally to this work.
[†] Work performed when Kun Liu worked as an intern in Deeplycurious.ai.
[1] Our code is available at https://github.com/liukun95/Prism-Module

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 5358–5362
Florence, Italy, July 28 - August 2, 2019.

to reduce noise from different aspects. As a side effect, the interpretability of models is improved since different abstract aspects could represent different semantics.

We introduce a method to train this module jointly with downstream model without extra training data. During training, the prism module learns to find the proper words to be replaced for each abstract aspect and also learns the embeddings of abstract aspects which can represent the task-related semantics of words. Furthermore, we introduce a novel trick to reduce the high variance in training brought by REINFORCE method.

The prism module can be easily integrated into downstream model to reduce noise and improve performance. We evaluate our method on NER task. Results show that our model outperforms the baseline by a substantial margin.

2 Related Work

Attention-based models achieve the state of the art performance in a broad range of NLP tasks. Although soft attention is more popular, hard attention is found to be more effective with good training (Xu et al., 2015). Hard attention has been successfully applied in computer vision (Ba et al., 2014; Mnih et al., 2014) but the application is limited in NLP. Lei et al. (2016) proposed a novel type of hard attention and apply it to improve the interpretability of models. However, the accuracy is not improved. Inspired by this, our proposed method can also be understood as hard-attention based but improves the accuracy successfully.

In addition to improving accuracy, attention-based models also improve the interpretability by showing the inner working of neural networks (Rush et al., 2015; Rocktäschel et al., 2015; Lei et al., 2016). Disentangling provides another way to improve the interpretability by extracting information from different aspects of the input. Lin et al. (2017) propose a multi-aspect self-attention to disentangle the latent semantic information of the input sentence. Jain et al. (2018) propose a model to learn disentangled representations of texts for 4 given biomedical aspects. Our proposed method can be regarded as the combination of the above two types of methods to improve the interpretability of the model.

3 Model

3.1 Prism Module

The target of this module is to get the sentences with less noise by replacing some of the words with abstract aspects. In a sentence, since each word has different semantics and contributes differently to the task, the key is to calculate the probability distribution over possible replacements.

Given a sentence X, which have n words

$$X = (w_1, w_2, w_3, \cdots w_n) \tag{1}$$

where w_i is the embedding of the i-th word in the sentence. We also have m different abstract aspects which represent m aspects of semantics

$$A = (a_1, a_2, a_3, \cdots a_m) \tag{2}$$

where a_i is the embedding of the i-th abstract aspect.

We apply bidirectional LSTM to the input sentence, which could capture some dependency between words.

$$\overrightarrow{h}_t = \overrightarrow{LSTM}\,(w_t, \overrightarrow{h}_{t-1}) \tag{3}$$

$$\overleftarrow{h}_t = \overleftarrow{LSTM}\,(w_t, \overleftarrow{h}_{t+1}) \tag{4}$$

where $\overrightarrow{h}_t$ and $\overleftarrow{h}_t$ denote the hidden states. We use h_t, the concatenation of $\overrightarrow{h}_t$ and $\overleftarrow{h}_t$ as the annotation of words. All n hidden states are annotated as the matrix

$$H = (h_1, h_2, h_3, \cdots h_n) \tag{5}$$

We define binary variable $s_{i,j} \in 0, 1$ which indicates whether j-th word w_j is replaced by i-th abstract aspect a_i or not. Then, the probabilities P with shape of m-by-n can be computed, where each element $p_{i,j}$ is the probability of $s_{i,j} = 1$. P is calculated as:

$$P = sigmoid(WH^T + b) \tag{6}$$

$$p_{i,j} = p(s_{i,j} = 1|X) \tag{7}$$

Here, W is the weight with the size of m-by-$2h$ and b is the bias.

$s_{i,j}$ is the random variable with multinoulli distribution parametrized by $p_{i,j}$. To get the replaced sentences, we sample S' according to the probability distribution $p_{i,j}$

$$S' = \begin{bmatrix} s'_{1,1} & \cdots & s'_{1,n} \\ \vdots & \ddots & \vdots \\ s'_{m,1} & \cdots & s'_{m,n} \end{bmatrix} \tag{8}$$

where i-th row of the matrix indicates which words in a sentence are replaced with i-th abstract aspect. After replacing the words with the guide of S', we obtain m replaced sentences $(X'_1, X'_2, X'_3 \cdots X'_m)$ where each one is denoised from different aspect. Then, these parallel sentences including m denoised sentences and the original sentence are used as the input of the downstream model.

3.2 Model Training

The prism module is trained jointly with downstream model. The parameters in the model can be divided into two parts, θ_o for downstream model and θ_a for prism module.

The objective for optimizing θ_o is to improve the prediction accuracy of the model. Since the input of the model includes both the word embeddings and abstract aspect embeddings, the loss function for parameters θ_o is

$$L(\theta_o) = L(\theta_o,\ X,\ y) + L(\theta_o,\ X,\ S',\ y) \tag{9}$$

The objective for optimizing θ_a is to replace proper words with proper abstract aspects. Because of the discrete variable $s_{i,j}$, the loss function is non-differentiable for the parameters θ_a. We use the policy gradient/REINFORCE (Williams, 1992) to optimize θ_a. Since we expect that not only the downstream model is well trained, but also the replaced sentences can achieve favorable performance in downstream task, the loss function $L(\theta_o)$ is used as reward R. The objective function for θ_a is:

$$L(\theta_a) = E_{s\sim p}(R \log(p(s|X))) \tag{10}$$

Besides, we also introduce a penalization term $\Omega(A)$ proposed by Lin et al. (2017) to diversify the abstract aspects which are expected to represent different disentangled aspects.

$$\Omega(A) = \left\| \widehat{A}\widehat{A}^T - I \right\|_F^2 \tag{11}$$

where $\|\|_F$ denotes the Frobenius norm of a matrix, I stands for the identity matrix and $\widehat{A}$ is calculated by normalizing each a_i of A.

Considering that we sample the S' according to the probability distribution to simplify the expectation, for all parameters, the loss function L is:

$$\begin{aligned} L &= L(\theta_o) + L(\theta_a) + \Omega(A) \\ &= L(\theta_o,\ X,\ y) + L(\theta_o,\ X,\ S',\ y) \\ &\quad + L(\theta_o) \log\left(p\left(S'|X\right)\right) + \Omega(A) \end{aligned} \tag{12}$$

3.3 Normalization of Reward

High variance is one of the disadvantages of REINFORCE method, which makes models difficult to converge. No exception, our model also suffers from the same problem. We propose a novel method to reduce the variance and stabilize the training process. We normalize the rewards by making them have the mean of 0 and variance of 1.

$$\mu \leftarrow \frac{1}{m}\sum_{i=1}^{m} R_i \tag{13}$$

$$\sigma^2 \leftarrow \frac{1}{m}\sum_{i=1}^{m} (R_i - \mu)^2 \tag{14}$$

$$\widehat{\mathrm{R_i}} \leftarrow \frac{R_i - \mu}{\sqrt{\sigma^2}} \tag{15}$$

where mean μ and variance σ are calculated over each mini-batch. $\widehat{\mathrm{R_i}}$ denotes the normalized reward. The loss L becomes

$$\begin{aligned} L &= L(\theta_o,\ X,\ y) + L(\theta_o,\ X,\ S',\ y) \\ &\quad + \widehat{\mathrm{R_i}}\ \log\left(p\left(S'|X\right)\right) + \Omega(A) \end{aligned} \tag{16}$$

4 Experiments

We evaluate the effectiveness of our noise reduction method on NER task.

Dataset: CoNLL 2003 (Sang and De Meulder, 2003) is used as our dataset.

Baseline: Yang et al. (2018) compare the performance of twelve neural sequence labeling models in NER task and the architecture CNN-BiLSTM (Bi-directional LSTM)-CRF (Ma and Hovy, 2016) achieves the best result (F1). Therefore, we use this model as our baseline.

Figure 1 shows our model where the prism module is integrated into CNN-BiLSTM-CRF architecture. The sentence is fed into the prism module and the output of this module is $m(e.g., 3)$ sentences which are denoised from different aspect. These $m + 1$ parallel sentences including the m denoised sentences and the original sentence are fed into BiSTM+CRF network to predict the labels. Besides, only the original sentence is used in testing.

4.1 Model Configuration

In the prism module, the hidden size of BiLSTM is the same as in CNN-BiLSTM-CRF architecture. The number of abstract aspects is set as 8. Except the hyper parameters in prism module, other hyper parameters are all set as (Ma and Hovy, 2016).

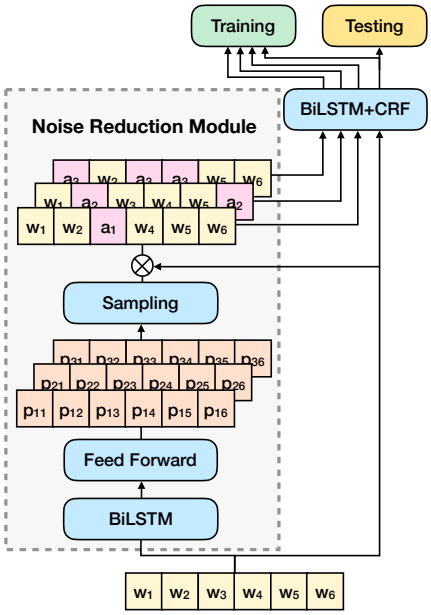

Figure 1: CNN-BiLSTM-CRF architecture with prism module. $w_1, w_2...$ denote the concatenation of original word embeddding and character-level representation which is computed by CNN.

Model	F1
Baseline (Ma and Hovy, 2016)	91.2
Multi-aspect hard attention	91.5
Random replacement	91.5
Single aspect	91.3
Our method	**91.8**

Table 1: NER F1 score of baseline, three ablation experiments and our model on test data of CoNLL-2003.

4.2 Result and Analysis

The experimental results are shown in Table 1. Our model outperforms the baseline by a clear margin.

To prove the effectiveness of our prism module, we design three ablation experiments:

Multi-aspect hard attention: Instead of replacing the words with abstract aspects, we replace the embeddings of selected words with zero vectors. This method can be regarded as a type of multi-aspect hard attention where some of the words are ignored.

Random replacement: Instead of learning to select the words to be replaced guided by the downstream task, we select the words to be replaced randomly for each abstract aspect. It is a kind of data noising technique which is similar to the method proposed in (Xie et al., 2017) with multiple aspects.

Figure 2: Heat map for S'

Single aspect: In our model, one word could be replaced with different abstract aspects in different denoised sentences. In this experiment, there is only one denoised sentence where each word could only be replaced with the abstract aspect of the maximum probability.

Our model has better performance than three ablation experiments as shown in Table 1. The results indicate that (1) The trainable embeddings of each abstract aspect can capture the information which is valuable for the task. (2) Our model can learn to replace words properly guided by the downstream task (e.g., NER). (3) For each word, more than one aspect of semantics are task-related. Additionally, considering that the first two ablation experiments improve F1 by 0.3% but the last one only improves 0.1%, multi-aspect denoising is important for the prism module.

4.3 Visualization

We visualize the matrix S' by drawing the heat map of each row vector as shown in Figure 2. In this example, *japan* and *china* are location entities. Each row corresponds to one abstract aspect and each element indicates whether this word is replaced. The heat map shows that each abstract aspect replaces some of words to keep certain task-related semantics and filter out other information. Since the abstract aspects represent different meanings respectively, the selections of words vary between rows which indicates noise is reduced from different aspects. From the heat map, we can also learn that a word can be replaced with multiple abstract aspects and this process is the disentanglement of semantics.

5 Conclusion

In this paper, we propose a prism module to reduce the noise of word embeddings by selectively replacing some words with task-related semantic aspects. We also introduce a structure to train this

prism module jointly with existing model and no extra data is needed. Considering REINFORCE method is used in training, a novel method is introduced to reduce the variance of rewards. As a result, our model outperforms the baseline by a clear margin and the ablation analysis proves the effectiveness of our method. As a side effect, this module also improves the interpretability of models. Since our prism module can be easily integrated into existing models, it can be applied in a wide range of neural architectures.

References

Jimmy Ba, Volodymyr Mnih, and Koray Kavukcuoglu. 2014. Multiple object recognition with visual attention. *arXiv preprint arXiv:1412.7755*.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155.

Sarthak Jain, Edward Banner, Jan-Willem van de Meent, Iain J Marshall, and Byron C Wallace. 2018. Learning disentangled representations of texts with application to biomedical abstracts. *arXiv preprint arXiv:1804.07212*.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. *arXiv preprint arXiv:1606.04155*.

Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130*.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. *arXiv preprint arXiv:1603.01354*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Volodymyr Mnih, Nicolas Heess, Alex Graves, et al. 2014. Recurrent models of visual attention. In *Advances in neural information processing systems*, pages 2204–2212.

Tim Rocktäschel, Edward Grefenstette, Karl Moritz Hermann, Tomáš Kočiskỳ, and Phil Blunsom. 2015. Reasoning about entailment with neural attention. *arXiv preprint arXiv:1509.06664*.

Alexander M Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. *arXiv preprint arXiv:1509.00685*.

Erik F Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050*.

Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256.

Ziang Xie, Sida I Wang, Jiwei Li, Daniel Lévy, Aiming Nie, Dan Jurafsky, and Andrew Y Ng. 2017. Data noising as smoothing in neural network language models. *arXiv preprint arXiv:1703.02573*.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057.

Jie Yang, Shuailong Liang, and Yue Zhang. 2018. Design challenges and misconceptions in neural sequence labeling. *arXiv preprint arXiv:1806.04470*.