Combining Word-Level and Character-Level Representations for Relation Classification of Informal Text

Dongyun Liang, Weiran Xu, Yinge Zhao,
PRIS, Beijing University of Posts and Telecommunications, China
{dongyunliang, xuweiran}@bupt.edu.cn yingezhao@outlook.com

Abstract

Word representation models have achieved great success in natural language processing tasks, such as relation classification. However, it does not always work on informal text, and the morphemes of some misspelling words may carry important short-distance semantic information. We propose a hybrid model, combining the merits of word-level and character-level representations to learn better representations on informal text. Experiments on two dataset of relation classification, SemEval-2010 Task8 and a large-scale one we compile from informal text, show that our model achieves a competitive result in the former and state-of-the-art with the other.

1 Introduction

Deep learning has made significant progress in natural language processing, and most of approaches treat word representations as the cornerstone. Though it is effective, word-level representation is inherently problematic: it assumes that each word type has its own vector that can vary independently; most words only occur once in training data and out-of-vocabulary(OOV) words cannot be addressed. A word may typically include a root and one or more affixes (*rock-s, red-ness, quick-ly, run-ning, un-expect-ed*), or more than one root in a compound (*black-board, rat-race*). It is reasonable to assume that words which share common components(root, prefix, suffix)may be potentially related, while word-level representation considers each word separately. On the other hand, new words enter English from every area of life, e.g. *Chillaxing* - Blend of *chilling* and *relaxing*, represent taking a break from stressful activities to rest or relax. Whereas the vocabulary size of word-level model is fixed beforehand, the lack of these word representations may lose important semantic information.

Especially on informal text, the problems of word-level representation will be amplified and hard to ignore. Recently, character-level representation, which takes characters as atomic units to derive the embeddings, demonstrates that it can memorize the arbitrary aspects of word orthography. Parameters of these simple model are less, and it will be not ideal when processing long sentence. Combining word-level and character-level representations attempts to overcome the weaknesses of the two representations.

We utilize a Bidirectional Gated Recurrent Unit (Bi-GRU) (Chung et al., 2014) and Convolutional Neural Networks(CNN) to capture two-level semantic representations respectively. While character-level information is likely to be drowned out by word-level information if simply connected, we adopt Highway Networks (Srivastava et al., 2015) to balance both. To evaluate our model, we evaluate on a public benchmark: SemEval-2010 Task8. This dataset is small and restricted in their relation types and their syntactic and lexical variations, and it is still unknown whether learning on the range of the specific relation transfers well to informal text. As such, we introduce a large-scale dataset based on the corpus and queries of TAC-KBP Slot Filling Track (Surdeanu and Ji, 2014) between 2009 to 2014, which contains 48k relation sentences, called KBP-SF48[1].

TAC-KBP corpus comes from newswire, Web, post and discussion forum documents actually comprised of informal content, including language mismatch and spelling errors. We extract sentences from slots and fillers of Slot Filling Evalua-

[1] https://github.com/waterblas/KBP-SF48

Proceedings of the 2nd Workshop on Representation Learning for NLP, pages 43–47,
Vancouver, Canada, August 3, 2017.

tion with position indicators to keep the same format as SemEval-2010 Task8. For instance, the following sentence with two nominals surrounded by position indicators belong to *org:founded_by* relation:

Bharara's office brought insider trading charges against <e1>Raj Rajaratnam <e1/>, the co-founder of hedge fund <e2>Galleon Group<e2/>.

2 Related Work

Some works (Mikolov et al., 2013; Pennington et al., 2014) started to learn semantic representations of word by unsupervised approaches. Recently, relation classification has focused on neural networks. Zeng et al. (2014) utilized CNN to learn patterns of relations from raw text data to make representative progress, but a potential problem is that CNN is not suitable for learning long-distance semantic information. Santos et al. (2015) proposed a similar model named CR-CNN, and replaced the cost function with a ranking-based function. Some models (Xu et al., 2015; Cai et al., 2016) leveraged the shortest dependency path(SDP) between two nominals. Others (Zhou et al., 2016; Wang et al., 2016) employed attention mechanism to capture more important semantic information.

Working to a new dataset KBP37, Zhang and Wang (2015) proposed a framework based on a bidirectional Recurrent Neural Network(RNN). However, all these methods depend on learning word-level distributed representation without utilizing morphological feature.

Recent work captures word orthography using character-based neural networks. dos Santos and Zadrozny (2014) proposed a deep neural network to learn character-level representation of words for POS Tagging. Zhang et al. (2015) demonstrated the effectiveness of character-level CNN in text classification. Kim et al. (2015) employed CNN and a highway network to learn rich semantic and orthographic features from encoding characters. There were some models (Ling et al., 2015; Dhingra et al., 2016) based on RNN structures, which can memorize arbitrary aspects of word orthography over characters.

Our model uses multi-channel GRU units and CNN architecture to learn the representations of word-level and character-level, and project it to a softmax output layer for relation classification.

3 Model

As shown in Figure 1, the model learns word-level and character-level representations respectively, and combines them with interaction to get the final representation.

3.1 Word-level

Given a relation sentence consisting of words $w_1, w_2, ..., w_m$, each w_i is defined as a one hot vector 1_{w_i}, with value 1 at index w_i and 0 in all other dimensionality. We multiply a matrix $P_W \in \mathbb{R}^{d_w \times |V|}$ by 1_{w_i} to project the word w_i into its word embedding x_i, as with a lookup table:

$$x_i = P_W w_i \tag{1}$$

where d_w is the size of word embedding and V is the vocabulary of training set.

Then input the $x_1, x_2, ..., x_m$ sequence to a Bi-GRU network iteratively. Each GRU unit apply the following transformations:

$$\begin{aligned} r_t &= \sigma(W_r x_t + U_r h_{t-1} + b_r) \\ z_t &= \sigma(W_z x_t + U_z h_{t-1} + b_z) \\ h_t &= (1 - z_t) \odot h_{t-1} + z_t \odot \widetilde{h}_t \\ \widetilde{h}_t &= \tanh(W_h x_t + U_h(r_t \odot h_{t-1}) + b_h) \end{aligned} \tag{2}$$

where z_t is a set of update gates, r_t is a set of reset gates and $\odot$ is an element-wise multiplication. W_r, W_z, W_h and U_r, U_z, U_h are weight matrices to be learned, and $\widetilde{h}_t$ is the candidate activation. We use element-wise sum to combine the forward and backward pass final states as word-level representation: $h_m^w = [\overrightarrow{h_m} + \overleftarrow{h_0}]$.

3.2 Character-level

To capture morphological features, we use convolutions to learn local n-gram features at the lower network layer. As character-level input, original sentence is decomposed into a sequence of characters, including special characters, such as white-space. We first project each character into a character embedding x_i by a lookup table whose mechanism is exactly as Eq.1.

Given the $x_1, x_2, , x_n$ embedding sequence, we compose the matrix $D^k \in \mathbb{R}^{kd_c \times n}$ to execute convolutions with same padding:

$$C^k = \tanh(W_{con}^k D^k) \tag{3}$$

where d_c is the size of word embedding and each column i in D^k consists of the concatenation of

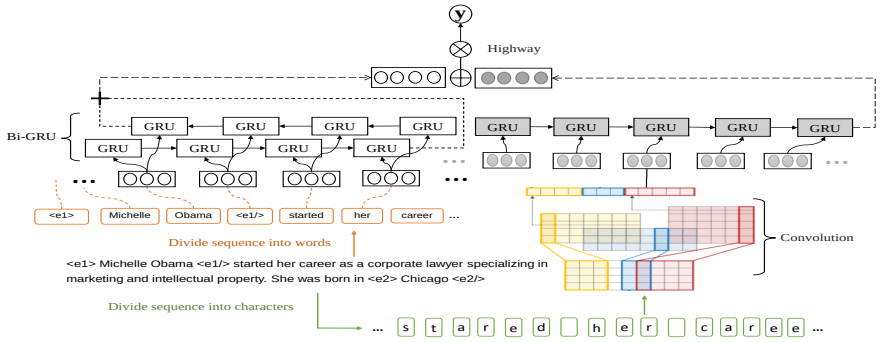

Figure 1: Hybrid model combining word-level and character-level representation.

vectors (i.e. k embeddings centered at the i-th character), W^k_{con} is a weight matrix of convolution layer, and $C^k \in \mathbb{R}^{c \times n}$ is the output of the convolution with c filters. We use p groups of filters with varying widths to obtain n-gram feature, and concatenate them by column:

$$C = C^{k_1} \oplus C^{k_2} \oplus ... \oplus C^{k_p} \quad (4)$$

The next step, $c_i, ..., c_n$ denoted by the column vector of C are fed as input sequence to a forward-GRU network(Eq.2), and we pick up final states activation h^c_n as character-level representation.

3.3 Combination

Instead of fully connected network layer, we utilize Highway Networks to emphasize impact of character level. Highway can be used to adaptively copy or transform representations, even when large depths are not required. We apply this idea to retain some independence of word and character when merging with interaction. Let h^* be the concatenation of h^w_m and h^c_n, The combination z is obtained by the Highway Network:

$$\begin{aligned} z &= t \odot g(W_H h^* + b_H) + (1 - t) \odot h^* \\ t &= \sigma(W_T h^* + b_T) \end{aligned} \quad (5)$$

where g is a nonlinear function ($\tanh$), t is referred to as the transform gate, and $(1 - t)$ as the carry gate. W_T and W_H are square weight matrices, and b_T and b_H are bias vectors.

3.4 Training

Training our model for classifying sentence relation is a processes to optimizing the whole parameters θ of network layers. Given a input sentence X and the candidate set of relation Y, the classifier returns output $\hat{y}$ as follows:

$$\hat{y} = \arg\max_{y \in Y} p(y|X, \theta) \quad (6)$$

We let the combination vector z through a softmax layer to give the distribution $y = softmax(W_f z + b_f)$.

The training objective is the penalized cross-entropy loss between predicted and true relation:

$$J(\theta) = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{m} t_{i,j} \log(y_{i,j}) + \lambda \|\theta\|^2_F \quad (7)$$

where N is the mini-batch size, m is the size of relation set, $\boldsymbol{t} \in \mathbb{R}^m$ denotes the one-hot represented ground truth, $y_{i,j}$ is the predicted probability that the $i-$th sentence belongs to class j, and λ is a coefficient of L2 regularization.

4 Experiments

4.1 Dataset

We evaluate our model on two dataset. SemEval-2010 Task8 dataset contains 9 directional relations and an Other class.

There exist dataset derived from TAC-KBP for relation classification, such as KBP37(20k example for evaluation) collected by (Zhang and Wang, 2015). Based on this and more public corpus of resent years, we introduce a new larger scale dataset, called KBP-SF48. There are 48,340 annotated examples distributed among 40 relations(excluding *no_relation* and *org:website*), including 33,838 sentences for training that consists of 102 unique characters, 9,668 for testing and 4,834 for validation.

Compared to SemEval-2010 Task8, the relation type of KBP-SF48 is designed to build a Knowledge Base from unstructured text, including quite a few informal documents, and the specific nominals that be-

longs to these relations can be filled in specific slots. There exists non-directional and the directional corresponding relations (e.g. per:children & per:parents and org:members & org:member_of).

4.2 Results

Model	F1
SVM (Rink and Harabagiu, 2010)	82.2
CNN (Zeng et al., 2014)	82.7
SDP-LSTM (Xu et al., 2015)	83.7
Att-BLSTM (Zhou et al., 2016)	84.0
BRCNN (Cai et al., 2016)	86.3
Ours	84.1

Table 1: Comparison on SemEval-2010 Task8.

Table 1 compares our model with other previous state-of-the-art methods on SemEval-2010 Task8 dataset. Rink and Harabagiu (2010) built a SVM classifier on a variety of handcrafted features, and achieved an F1-score of 82.2%. Xu et al. (2015) achieved an F1-score of 83.7% via heterogeneous information along the SDP. BRCNN (Cai et al., 2016) combined CNN and two-channel LSTM units to learns features along SDP, and made use of POS tags, NER and WordNet hypernyms. Att-BLSTM (Zhou et al., 2016) only operated attention mechanism on Bidirectional Long Short-Term Memory(BLSTM) units with word vector.

Our model yields an F1-score of 84.1%, and outperforms most of the existing competing approaches without using any human-designed features and lexical resources.

On KBP-SF48 benchmark, we evaluate our model by top 1 precision, and mean rank of correct relation because of the existence of non-directional relations,

We reproduce the results on our own to show the performances of the other systems with the same train/dev/test splits, and ablate different aspects of the proposed model to show the impact of every component of our architecture. As is seen from Table 2, our model achieves a state-of-the-art result on KBP-SF48 dataset. Our model has already outperformed the RNN-based (Zhang and Wang, 2015) model of the KBP37 dataset,

Model	Precision @1	Mean Rank
RNN-based (Zhang and Wang, 2015)	68.9%	2.01
CNN (Zeng et al., 2014)	79.1%	1.55
BLSTM and Att-BLSTM (Zhou et al., 2016)	78.9% 80.2%	1.59 1.51
Character-level Only (Dhingra et al., 2016)	74.9%	1.85
Word-level Only	78.4%	1.60
Full connected network	80.9%	1.51
Ours	81.7%	1.45

Table 2: Comparison on KBP-SF48

a small scale dataset based on TAC-KBP Slot Filling Track. We compare our results against some state-of-the-art methods (Zeng et al., 2014; Zhou et al., 2016) of SemEval-2010 Task8, and our model achieves a better result by combining character feature into word-level representation. Then, we illustrate Bi-GRU architecture of Tweet2Vec (Dhingra et al., 2016), a pure character-level composition model, to show the effectiveness of character-level representation. Next, we get rid of the impact of characters to do word-level only experiment, and replace the highway with a fully connected layer. These clean comparisons demonstrate that the character-level and Highway network help to learn a better representation for classification.

5 Conclusion

In this paper, we propose a hybrid model that combines word-level and character-level representations. This model encodes characters by a cascade of CNN and GRU units, encodes words by Bi-GRU units, and uses Highway Network to combine. We demonstrate that our model achieves competitive results on the popular benchmark SemEval-2010 Task8 and achieves a better performance at learning character features on the KBP-SF48 dataset without relying on any lexical resources. In future, we plan to add interactions for each word with the corresponding positional characters.

References

Rui Cai, Xiaodong Zhang, and Houfeng Wang. 2016. Bidirectional recurrent convolutional neural network for relation classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. pages 756–765.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555* .

Bhuwan Dhingra, Zhong Zhou, Dylan Fitzpatrick, Michael Muehl, and William W Cohen. 2016. Tweet2vec: Character-based distributed representations for social media. *arXiv preprint arXiv:1605.03481* .

Cícero Nogueira dos Santos and Bianca Zadrozny. 2014. Learning character-level representations for part-of-speech tagging. In *ICML*. pages 1818–1826.

Yoon Kim, Yacine Jernite, David Sontag, and Alexander M Rush. 2015. Character-aware neural language models. *arXiv preprint arXiv:1508.06615* .

Wang Ling, Tiago Luís, Luís Marujo, Ramón Fernandez Astudillo, Silvio Amir, Chris Dyer, Alan W Black, and Isabel Trancoso. 2015. Finding function in form: Compositional character models for open vocabulary word representation. *arXiv preprint arXiv:1508.02096* .

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

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*. volume 14, pages 1532–43.

Bryan Rink and Sanda Harabagiu. 2010. Utd: Classifying semantic relations by combining lexical and semantic resources. *ACL 2010* page 256.

Cicero Nogueira dos Santos, Bing Xiang, and Bowen Zhou. 2015. Classifying relations by ranking with convolutional neural networks. *arXiv preprint arXiv:1504.06580* .

Rupesh K Srivastava, Klaus Greff, and Jürgen Schmidhuber. 2015. Training very deep networks. In *Advances in neural information processing systems*. pages 2377–2385.

Mihai Surdeanu and Heng Ji. 2014. Overview of the english slot filling track at the tac2014 knowledge base population evaluation. In *Proc. Text Analysis Conference (TAC2014)*.

Linlin Wang, Zhu Cao, Gerard de Melo, and Zhiyuan Liu. 2016. Relation classification via multi-level attention cnns. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. pages 1298–1307.

Yan Xu, Lili Mou, Ge Li, Yunchuan Chen, Hao Peng, and Zhi Jin. 2015. Classifying relations via long short term memory networks along shortest dependency paths. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*.

Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, Jun Zhao, et al. 2014. Relation classification via convolutional deep neural network. In *COLING*. pages 2335–2344.

Dongxu Zhang and Dong Wang. 2015. Relation classification via recurrent neural network. *arXiv preprint arXiv:1508.01006* .

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*. pages 649–657.

Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016. Attention-based bidirectional long short-term memory networks for relation classification. In *The 54th Annual Meeting of the Association for Computational Linguistics*. page 207.