Improve Neural Mention Detection and Classification via Enforced Training and Inference Consistency

Huasha Zhao[1], Yi Yang[1,2] *, Qiong Zhang[1], and Luo Si[1]

[1]Alibaba Group, San Mateo, CA
{huasha.zhao, qz.zhang, luo.si}@alibaba-inc.com
[2]Nanjing University, Nanjing, Jiangsu, China
yangyi868@gmail.com

Abstract

This is our first time participating the KBP tracks. We focus on the English mention detection and classification tasks this year. Our system achieves F_1 score of 0.811 in 2017 English NERC evaluation, which ranks the first among all participants. The system is developed based on a widely adopted BiLSTM-CRF model which is considered as state of the art for many sequence labeling tasks. In this paper, we apply a multi-task version of BiLSTM-CRF model to the NERC task, to better utilize additional data sources. Furthermore, novel methods are proposed to enforce data and label consistency at both training and prediction time. Extensive experiments show that our methods significantly improve the performance over the baseline model.

1 Introduction

Mention detection (MD) and classification are fundamental tasks in Natural Language Processing (NLP). Mention classification are more widely know as Entity Recognition(ER) which includes named entity recognition and nominal entity recognition. MD and ER are the building blocks for higher level applications such as natural language understanding, machine reading, etc. They are usually treated as sequence labeling problems. Although the topics have been studied extensively for the past several decades, development of neural network and deep learning based methods in recent years (Lample et al., 2016; Ma and Hovy, 2016; Yang et al., 2017; Kenton Lee and Zettlemoyer, 2017; Xinchi Chen, 2017) significantly improve the previous state-of-the-art.

A popular neural architecture for mention detection and classification is BiLSTM-CRF (Lample et al., 2016). The architecture has been shown to achieve bests performance on many sequence labeling problems. In real world applications, training data in the specific domain of interest is usually not enough to achieve best performance. As a result, external data is needed to improve model performance. For example, in the case of KBP 2016 tracks, both the 1*st* and the 2*nd* teams (ranking in the NERC evaluation) use external annotations (Liu et al., 2016; Xu et al., 2017). In most cases, data distribution and labeling guideline from source domain (external data) is different from that of target domain (domain of interest). Therefore, training a model by simply combining the target and source data may not yield satisfactory results. Fortunately, BiLSTM-CRF architecture provides a natural way to model the heterogeneity of the training data.

In this work, we apply a multi-task BiLSTM-CRF architecture to the mention detection and classification problem, with additional entity type embeddings and domain adaption. Two novel methods based on the theme of *consistency* are proposed to improve the model performance.

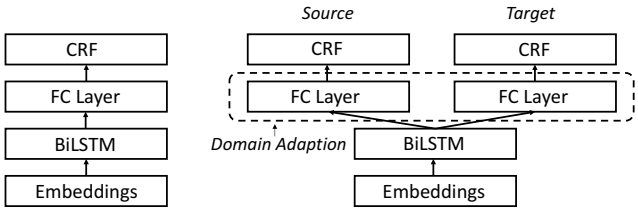
Training Data Consistency

To ensure homogeneity between source and target training data, adaptive training data selection is applied to source data to filter out instances with misaligned annotation guideline. Data selection is interleaved with model training iteratively, and this training process terminates until convergence.

Prediction Label Consistency

Global label consistency is enforced at prediction time. The goal is to capture document level contexts. A phrase is likely to be an entity if it is detected in another sentence in the same document. It also helps detect related mentions, such as the mention *jobs* is more likely to be a PER when it occurs in the same discussion forum with *Apple*.

* Work was done while doing internship at Alibaba.



(a) (b)

Figure 1: Neural architectures for mention detection and classification. a) Single-task model. b) Multi-task model with domain adaptions.

2 Related Works

There are many works in literature on applying neural networks to mention detection and entity recognition problems (Lample et al., 2016; Ma and Hovy, 2016; Yang et al., 2017; Peng and Dredze, 2016). Our work is most closed to (Yang et al., 2017). However, we introduce additional channel in the embedding layer and add domain adaption layers (Peng and Dredze, 2016).

The idea of training data consistency is derived from topics of data selection (Moore and Lewis, 2010) and instance weighting (Jiang and Zhai, 2007) from the transfer learning community. Different from previous work, we propose an adaptive selection approach interleaved with BiLSTM-CRF model training. Global label consistency has been studied in (Radford et al., 2015; Krishnan and Manning, 2006). Here we share similar ideas with previous work, but explore the use of a recent neural network architecture, i.e. a dilated CNN (Strubell et al., 2017), to the task, and compare with dictionary based method.

3 Approach

This section describes the model used for the KBP task. We first describe a slight variant of BiLSTM and it's multi-task version for transfer learning. Then we present in details how the theme of *consistency* is applied to further improve the performance.

3.1 BiLSTM-CRF

BiLSTM-CRF is a widely adopted neural architecture for sequence labeling problems including MD and ER. BiLSTM-CRF is a hierarchical model and the architecture is illustrated in Figure 1(a).

The first layer of the model maps words to their embeddings. Let $\mathbf{x} = (x_1, \cdots, x_n)$ denote a sentence composed of n words in a sequence, with $x_i's$ as their word embeddings. In the second layer, word embeddings are encoded using a bidirectional-LSTM network, and the output is $\mathbf{h} = (h_1, \cdots, h_n)$, where $h_t = BiLSTM(\mathbf{x}, t)$. The encodings are further passed to a fully connection network, to compute CRF features $\phi(\mathbf{x}) = G \cdot \mathbf{h}$, and finally objective to optimize is the likelihood defined as the following,

$$p(\mathbf{y}|\mathbf{x};\theta) = \frac{\prod_{i=1}^{n} \exp(\theta \cdot f(y_{i-1}, y_i, \phi(\mathbf{x})))}{Z}, \tag{1}$$

where $\mathbf{y}$ are predicted labels and Z is the normalizing constant.

3.1.1 Character and Entity Embeddings

We extend the vanilla BiLSTM-CRF model by adding character and entity embeddings to the embedding layer. x_i is the concatenation of word embeddings, character embeddings and its entity embeddings, $x_i = [\omega_i, c_i, g_i]$. Character embeddings are modeled using another bidirectional LSTM network at character level. Entity embeddings are derived from a noisy gazetteer created using wikipedia articles. The gazetteer is derived from the word-entity statistics from (Pan et al., 2017). More specifically, each coordinate of the entity embedding is the probability of a word occurring as the corresponding entity type.

3.1.2 Domain Adaption

To explore external datasets, we apply multi-task (MT) BiLSTM-CRF with domain adaptions, as illustrated in Figure 1(b). The fully connection layer are adapted to different datasets. The CRF features are computed separately, i.e. $\phi^{\mathcal{T}}(\mathbf{x}) = G^{\mathcal{T}} \cdot \mathbf{h}$, $\phi^{\mathcal{S}}(\mathbf{x}) = G^{\mathcal{S}} \cdot \mathbf{h}$ for target and source dataset respectively. The loss function $p(\mathbf{y}|\mathbf{x};\theta^{\mathcal{T}})$, $p(\mathbf{y}|\mathbf{x};\theta^{\mathcal{S}})$ are optimized in alternating order.

3.2 Adaptive Training Data Consistency

Multi-task training can alleviate some of the problem caused by data heterogeneity between target and source. This section presents an adaptive data selection algorithm that further removes noisy data from source dataset.

The data selection procedure is described in details in Algorithm 1. At each iteration, data selection from the source domain is interleaved with model parameter updates. Training data is selected based on a *consistency score*, which measures the similarity between target and source data

Algorithm 1 Adaptive Training Data Selection

Input: Target training dataset $(\mathbf{x}, \mathbf{y}) \in \mathcal{T}$, source training dataset $(\mathbf{x}', \mathbf{y}') \in \mathcal{S}$.

Initialize: $\mathcal{S}_{train} \leftarrow \mathcal{S}$; $\mathcal{X}^S = \{\mathbf{x}' : (\mathbf{x}', \mathbf{y}') \in \mathcal{S}\}$.

Repeat:

1. Train the model for one iteration, by optimizing the following instance weighted object function,

$$J = \sum_{(\mathbf{x},\mathbf{y})\in\mathcal{T}} p(\mathbf{y}|\mathbf{x};\theta^{\mathcal{T}}) + \sum_{(\mathbf{x}',\mathbf{y}')\in\mathcal{S}train} p(\mathbf{y}'|\mathbf{x}';\theta^{\mathcal{S}}),$$

2. Compute consistency score for each training example in $\mathcal{S}$,

$$s(\mathbf{x}) = \max_i \sum_j p(x_i = j) \log \frac{p(x_i = j)}{q(x_i = j)},$$

   where $p(x_i) \sim softmax(\phi^{\mathcal{T}}(x_i))$ and $q(x_i) \sim softmax(\phi^{\mathcal{S}}(x_i)))$;

3. Construct $\mathcal{S}_{same}$, $\mathcal{S}_{diff}$ by the following, $\mathcal{S}_{same} = \{\mathbf{x} \in \mathcal{X}^S : s(\mathbf{x}) < \alpha\}$ and $\mathcal{S}_{diff} = \{\mathbf{x} \in \mathcal{X}^S : s(\mathbf{x}) > \beta\}$;

4. Update source training set $\mathcal{S}_{train}$, $\mathcal{S}_{train} \leftarrow \mathcal{S}_{train} \cup \mathcal{S}_{same} \setminus \mathcal{S}_{diff}$.

Until: $|S_{diff}| < k$

Return: the final BiLSTM-CRF model.

distribution. Specifically, the consistency score is derived from the KL divergence between $\phi^{\mathcal{T}}(\mathbf{x})$ and $\phi^{\mathcal{S}}(\mathbf{x})$ for every word in the sentence in the source training data. The iterations terminate until there is few additional data to filter out , up to a manually-set threshold.

3.3 Prediction Label Consistency

Arguably, a document level BiLSTM-CRF can help capture the global consistency and context, however, our experiment shows that document level model underperform the sentence level model if trained in one shot. This is not a surprise due to the low memory capacity of RNN models. As a result to a second pass approach for the problem. We experiment two approaches for enforcing label consistency.

3.3.1 Dictionary-based label consistency

The dictionary-based approach maintains a dictionary of all predicted named entities for each document, and all entities in the dictionary are enforced to the sentences where the entities are not recognized in a second pass of the prediction. We resolve classification conflicts in a random fashion.

3.3.2 Model-based label consistency

In a model-based approach, we use word embeddings and predicted labels as model inputs, and outputs are ground truth labels. The models are trained at document level to learn global label consistency. Dilated CNN is applied to model the inputs/outputs relationship. We choose CNN models because they are faster to train and have better memory capacities for longer texts, which overcomes the shortcomings of RNN models.

4 Experiments

This section presents experiments results of our methods on the KBP evaluation datasets. We focus on Engilsh mention detection and classification, which include both named entity recognition (NAM) and nominal entity recognition (NOM). The neural models are implemented using Tensorflow (Abadi et al., 2016). Dropout and gradient clipping are applied when necessary to avoid numerical issues during training. Performance numbers are reported using the `NERC` score as defined in (Ji et al., 2016).

4.1 Datasets

KBP 2015 data is used for evaluation on the 2016 evaluation dataset. Both datasets are used for training for KBP 2017 evaluation. We also leverage external data sources to improve model performance. Unlike (Liu et al., 2016), manual annotation is not feasible to us due to budget limit, we instead use ACE (Walker et al., 2006) and ERE (Song et al., 2015) entity annotations as source datasets. It is worth noting that annotation guidelines are different from one dataset to another, especially for nominal entity annotations.

4.2 Baseline

The baseline is a BiLSTM-CRF model with only word and character embeddings with source and target data combined as training data. GloVe vectors are used as word embeddings. NAM and NOM models are trained separately with individually tuned parameters.

Methods	NAM	NOM	Overall
baseline	0.809	0.587	0.748
+ entity embeddings	0.842	0.587	0.770

Table 1: Effectiveness of additional entity embeddings in model embedding layer.

Methods	NAM	NOM	Overall
baseline + entity embeddings	0.842	0.587	0.770
+MT	0.841	0.626	0.786
+MT + adaptive data selection	0.842	0.634	0.788

Table 2: Effectiveness of training data consistency.

4.3 Results

First, we examine the performance impact of entity embedding. As shown in Table 1, entity embedding is very useful for both NAM and NOM prediction tasks, and provides an overall performance improvement of 2.2 F_1 points. The entity embeddings are derived from soft gazetteer features. This experiment confirms the usefulness of gazetteer even in neural network models.

Next the effectiveness of training data consistency is evaluated. We compare MT domain adapted models and adaptive data selection with the baseline. Results in Table 2 show that both MT and adaptive data selection can significantly improve NOM detection. However, there is no gain at all for NAM detection. We manually evaluate the source and target datasets, and find that the annotation guideline and data distribution of NAM data are quite the similar while there are some significant differences for NOM data. Notably, many of the plural form nouns are marked as nominal entities in the ACE dataset while in our target KBP tasks plural nouns are not entities in general.

Table 3 presents the performance impact of prediction label consistency. Both dictionary and model based approaches improve the overall F1 score. Dictionary based approach does not change the NOM performance because only named entities are included in the dictionary. Optimal performance of model based methods is obtained using dilated CNN with 4 layers (with effective context size of 31) (Strubell et al., 2017). Wider context window does not improve model performance due to sparsity of data.

Methods	NAM	NOM	Overall
baseline + entity embeddings	0.842	0.587	0.770
+ label consistency (dictionary based)	0.851	0.587	0.778
+ label consistency (model based)	0.850	0.595	0.779

Table 3: Effectiveness of prediction label consistency.

Ensemble config	Precision	Recall	F1
Single model	0.833	0.760	0.795
2/4 voting	0.827	0.790	0.808
3/4 voting	0.850	0.776	0.811
Union of two 2/4	0.831	0.791	0.811

Table 4: Overall F1 score with different ensemble configurations.

The final model we submitted to the KBP track are ensembles. We experiment different ensemble configurations and the results are shown in Table 4. m/n voting means a voting based ensemble approach which selects a prediction if it's produced by m out of n models. There is a clear precision and recall trade-off between 2/4 and 3/4 voting. The model is more stable with 3/4 voting or two 2/4 voting combined.

Finally, we presents the evaluation results on both 2016 and 2017 datasets, and compare them with best scores of all KBP participants. As we can see from Table 5, the additional training data for KBP 2017 increases the overall model performance by 0.7 F_1 points.

Year	Our F1	Best F1
2016	0.804	0.772
2017	0.811	0.811

Table 5: Performance comparison between 2016 and 2017 datasets.

5 Conclusion and Future Works

This paper presents novel methods to improve neural mention detection and classification tasks, based on a theme of *consistency*. Extensive experiments show the effectiveness of the methods. Work needs to be done to justify in theoretic foundation the adaptive data selection algorithm. We

also plan to apply the methods described in this paper to other languages.

References

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. 2016. Tensorflow: A system for large-scale machine learning. In *OSDI*. volume 16, pages 265–283.

Heng Ji, Joel Nothman, Hoa Trang Dang, and Sydney Informatics Hub. 2016. Overview of tac-kbp2016 tri-lingual edl and its impact on end-to-end cold-start kbp. *Proceedings of TAC* .

Jing Jiang and ChengXiang Zhai. 2007. Instance weighting for domain adaptation in nlp. In *ACL*. volume 7, pages 264–271.

Luheng He Mike Lewis Kenton Lee and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In *EMNLP*. pages 188–197.

Vijay Krishnan and Christopher D Manning. 2006. An effective two-stage model for exploiting non-local dependencies in named entity recognition. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pages 1121–1128.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360* .

Dan Liu, Wei Lin, Shiliang Zhang, Si Wei, and Hui Jiang. 2016. The ustc nelslip systems for trilingual entity detection and linking tasks at tac kbp 2016 .

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

Robert C Moore and William Lewis. 2010. Intelligent selection of language model training data. In *Proceedings of the ACL 2010 conference short papers*. Association for Computational Linguistics, pages 220–224.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *ACL*.

Nanyun Peng and Mark Dredze. 2016. Multi-task multi-domain representation learning for sequence tagging. *arXiv preprint arXiv:1608.02689* .

Will Radford, Xavier Carreras, and James Henderson. 2015. Named entity recognition with document-specific kb tag gazetteers. In *EMNLP*. pages 512–517.

Zhiyi Song, Ann Bies, Stephanie Strassel, Tom Riese, Justin Mott, Joe Ellis, Jonathan Wright, Seth Kulick, Neville Ryant, and Xiaoyi Ma. 2015. From light to rich ere: annotation of entities, relations, and events. In *Proceedings of the 3rd Workshop on EVENTS at the NAACL-HLT*. pages 89–98.

Emma Strubell, Patrick Verga, David Belanger, and Andrew McCallum. 2017. Fast and accurate entity recognition with iterated dilated convolutions. In *EMNLP*.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. Ace 2005 multilingual training corpus. *Linguistic Data Consortium, Philadelphia* 57.

Zhan Shi Xipeng Qiu Xuanjing Huang Xinchi Chen. 2017. Adversarial multi-criteria learning for chinese word segmentation. In *ACL*.

Mingbin Xu, Hui Jiang, and Sedtawut Watcharawittayakul. 2017. A local detection approach for named entity recognition and mention detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. volume 1, pages 1237–1247.

Zhilin Yang, Ruslan Salakhutdinov, and William W Cohen. 2017. Transfer learning for sequence tagging with hierarchical recurrent networks. *ICLR* .