Improve Neural Entity Recognition via Multi-Task Data Selection and Constrained Decoding

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

Entity recognition is a widely benchmarked task in natural language processing due to its massive applications. The state-of-the-art solution applies a neural architecture named BiLSTM-CRF to model the language sequences. In this paper, we propose an entity recognition system that improves this neural architecture with two novel techniques. The first technique is Multi-Task Data Selection, which ensures the consistency of data distribution and labeling guidelines between source and target datasets. The other one is constrained decoding using knowledge base. The decoder of the model operates at the document level, and leverages global and external information sources to further improve performance. Extensive experiments have been conducted to show the advantages of each technique. Our system achieves state-of-the-art results on the English entity recognition task in KBP 2017 official evaluation, and it also yields very strong results in other languages.

1 Introduction

Entity Recognition (ER) is a fundamental task in Natural Language Processing (NLP). The task includes named entity recognition and nominal entity recognition. ER is the building blocks for higher level applications such as natural language understanding, question answering, machine reading comprehension, 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 improves the previous state-of-the-art.

A popular neural architecture for ER is BiLSTM-CRF (Lample et al., 2016). The architecture has been shown to achieve best performance on many sequence labeling tasks. In addition, the architecture can be easily extended to model different sources of training data. In real world applications, it is important to include external data sources for model training, because using only domain-specific data for training is usually not enough to achieve best 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 data source (Liu et al., 2016; Xu et al., 2017) for model training. The challenge here is to transfer knowledge from external data source to target data source. Multi-Task (MT) BiLSTM-CRF architecture (Yang et al., 2017) is designed for this knowledge transfer.

In this work, we develop an ER model based on the MT BiLSTM-CRF architecture, with additional entity embeddings and domain adaption. Two novel methods are proposed to further improve the model performance.

Multi-Task Data Selection

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

Constrained Decoding using Knowledge Base

Knowledge-based constraints are enforced at decoding time. The goal is to capture document level contexts given those knowledge. For example, 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

* Work was done while doing internship at Alibaba.

Proceedings of NAACL-HLT 2018, pages 346–351
New Orleans, Louisiana, June 1 - 6, 2018.

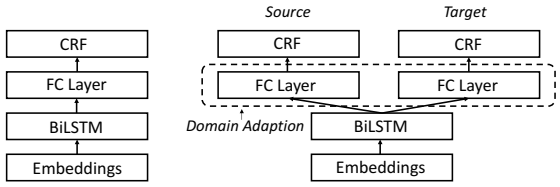

Figure 1: Neural architectures for mention detection and classification. a) Single-task model. b) Multi-task model with domain adaptions.

apple is more likely to be a ORG when it occurs in the same discussion forum with *Apple Inc*.

2 Related Works

There are many works in literature applying neural networks to ER problems (Lample et al., 2016; Ma and Hovy, 2016; Yang et al., 2017; Peng and Dredze, 2016). The baseline model of this work is mostly closed to (Yang et al., 2017). However, we introduce additional channel in the embedding layer(Peng and Dredze, 2016).

The idea of multi-task data selection 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 MT BiLSTM-CRF model training. Decoding with global constraints has been studied in (Yarowsky, 1993; Krishnan and Manning, 2006). Here we share similar ideas with previous work, but explore the use of external knowledge base (Radford et al., 2015) as constraints.

3 Approach

This section describes the baseline model used for the ER task. We first describe a slight variant of BiLSTM-CRF and its MT version for transfer learning. For the sake of brevity, discussions of the basis theory of MT learning are skipped and more details can be found in (Zhang and Yang, 2017). Then we present in details how data selection and constrained decoding are applied to further improve the model performance.

3.1 BiLSTM-CRF

BiLSTM-CRF is a widely adopted neural architecture for sequence labeling problems including 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/character embedding combinations. 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 CRF likelihood defined as the following,

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

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

3.1.1 Entity Embeddings

We extend the BiLSTM-CRF model by adding entity embedding channel to the embedding layer. As a result, x_i is the concatenation of word embedding, character embedding and its entity embedding, $x_i = [\omega_i, c_i, g_i]$. 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 distribution of a word occurring as the corresponding entity type.

3.1.2 Domain Adaption

To explore external datasets, we apply 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 $p(\mathbf{y}|\mathbf{x};\theta^{\mathcal{T}})$ and $p(\mathbf{y}|\mathbf{x};\theta^{\mathcal{S}})$ are optimized in alternating order.

3.2 Multi-task Data Selection

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 during multi-task training 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

Algorithm 1 Multi-task 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 $\mathcal{S}_{diff} = \{\mathbf{x} \in \mathcal{X}^S : s(\mathbf{x}) > \beta\}$; Thresholds α and β are manually set that determine the selection/exclusion of a data point.

4. Update source training set $\mathcal{S}_{train}$, $\mathcal{S}_{train} \leftarrow \mathcal{S}_{train} \cup \mathcal{S}_{same} \setminus \mathcal{S}_{diff}$. In the new training set, data with different distributions are eliminated.

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

Return: the final BiLSTM-CRF model.

based on a *consistency score*, which measures the similarity between target and source data 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. According to step 4, data that are not *consistent* with the target are eliminated from the training dataset. The iterations terminate until there is few additional data to filter out, up to a manually-tuned threshold.

3.3 Constrained Decoding using Knowledge Base

It has been well studied that non-local information can be used to help improve entity recognition performance (Radford et al., 2015) (Krishnan and Manning, 2006). Here we describe a globally constrained decoding (Graves et al., 2012) method used in our model. In particular, we use external knowledge information to guide the decoding process at the document level.

3.3.1 Knowledge Base

An external knowledge base is built from Wikipedia articles (Radford et al., 2015) (Dalton et al., 2014). For each Wikipedia entity, we first extract all its aliases from the redirects, and then build a cluster of the mentions for the this entity which includes all its aliases. Our goal is that given a document mentions *Microsoft*, the knowledge base can help identify the other mentions such as *MS Corp*. The knowledge base can be naturally extended to include related entities (using anchor texts), instead of only aliases of the same entity, in the cluster; we leave this to the future works.

Then we apply global decoding with constraint C, such that all mentions that belong to the same cluster should be labeled as the same entity type within a single document,

$$\mathbf{y_{1:N}} = \underset{C}{\operatorname{argmax}}\, p(\mathbf{y_{1:N}}|\mathbf{x_{1:N}};\theta),$$

where subscripts $1:N$ are indices of sentences within the same document. We use a greedy algorithm for decoding.

4 Experiments

This section presents experiments results of our methods on the KBP 2016 and 2017 evaluation datasets. We focus on Engilsh (ENG) and Mandarin Chinese (CMN) ER tasks, 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 F_1 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

Method	NAM	NOM	Overall
baseline (ENG)	0.809	0.587	0.748
+ EE (ENG)	0.842	0.587	0.770
baseline (CMN)	0.822	0.305	0.727
+ EE (CMN)	0.851	0.305	0.752

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

(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 word and character embeddings which simply combines source and target data as training data. GloVe vectors (Pennington et al., 2014) are used as word embeddings. NAM and NOM models are trained separately with individually tuned parameters.

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 for both languages. It provides an overall performance improvement of 2.2 F_1 points. Since the entity embeddings are derived from soft gazetteer features, this experiment confirms again the usefulness of gazetteer even in neural network models. In theory, the entity embeddings should have been already captured by the model itself; the additional predictability of the entity embeddings actually comes from the external dataset (Wikipedia) where the embeddings are derived from.

Next the effectiveness of Multi-Task Data Selection is evaluated. Results in Table 2 show that both MT and MTDS can significantly improve NOM detection over the baseline, and adaptive data selection in MTDS further improves over the MT model. However, there is no gain at all for NAM detection for both languages. 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 labeled as entities in general.

Method	NAM	NOM	Overall
baseline+EE (ENG)	0.842	0.587	0.770
+MT (ENG)	0.842	0.626	0.786
+MTDS (ENG)	0.842	0.634	0.788
baseline+EE (CMN)	0.851	0.305	0.752
+MT (CMN)	0.851	0.351	0.756
+MTDS (CMN)	0.851	0.364	0.758

Table 2: Effectiveness of Multi-Task Data Selection (MTDS).

Table 3 presents the performance impact of knowledge based constrained decoding. It is worth noting that the performance gain in the Chinese language is more limited in comparison with English. The primary reason behind this is that the English Wikipedia site is more comprehensive than its Chinese counterpart. Constrained decoding does not change the NOM performance because only name mentions are included in the knowledge base.

Method	NAM	NOM	Overall
baseline+EE (ENG)	0.842	0.587	0.770
+CD (ENG)	0.851	0.587	0.778
baseline+EE (CMN)	0.851	0.305	0.752
+CD (CMN)	0.855	0.305	0.754

Table 3: Effectiveness of Constrained Decoding (CD) using Knowledge Base.

Finally, we use model ensemble to further improve model scores. Four models are combined together for final evaluation. Majority vote is applied to produce final results. We presents the evaluation results on both KBP 2016 and 2017 datasets in Table 4, and compare them with state-of-the-art scores (Ji et al., 2016) (Ji et al., 2017). Our system ranks 1st in the English entity recognition task in the official evaluation in 2017. We also perform very strongly in the Chinese language as well: the best team applies many hand-tuned rules in the evaluation (Ji et al., 2017), while our model is free of rules. It also can be concluded from the table that the additional training data for KBP 2016 increases the overall model performance by 0.7 F_1 points.

Year/Language	Our F1	Best F1
2016/ENG	0.804	0.772
2017/ENG (Official evaluation)	0.811	0.811
2017/CMN	0.769	0.780

Table 4: Performance comparison between 2016 and 2017 datasets.

5 Conclusion and Future Works

This paper presents novel methods to improve neural entity recognition tasks. Multi-task data selection removes noise from training data, while constrained decoding further improves the model by exploiting global and external information sources. Extensive experiments show the effectiveness of the methods. Work needs to be done to justify in theoretic foundation the adaptive data selection algorithm. Furthermore, runtime and computational complexity of the system should be studied. We also plan to extend the knowledge base cluster to include related entities.

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.

Jeffrey Dalton, Laura Dietz, and James Allan. 2014. Entity query feature expansion using knowledge base links. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. ACM, pages 365–374.

Alex Graves et al. 2012. *Supervised sequence labelling with recurrent neural networks*, volume 385. Springer.

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* .

Heng Ji, Xiaoman Pan, Boliang Zhang, Joel Nothman, James Mayfield, Paul McNamee, and Cash Costello. 2017. Overview of tac-kbp2017 13 languages entity discovery and linking. *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* .

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. pages 1532–1543.

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.

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* .

David Yarowsky. 1993. One sense per collocation. Technical report, PENNSYLVANIA UNIV PHILADELPHIA DEPT OF COMPUTER AND INFORMATION SCIENCE.

Yu Zhang and Qiang Yang. 2017. A survey on multi-task learning. *arXiv preprint arXiv:1707.08114* .