Supplementary Material for “Better Modeling of Incomplete Annotations for Named Entity Recognition”

Zhanming Jie[1], **Pengjun Xie**[2], **Wei Lu**[1], **Ruixue Ding**[2] and **Linlin Li**[2]
[1]StatNLP Research Group, Singapore University of Technology and Design
[2]DAMO Academy, Alibaba Group
`zhanming_jie@mymail.sutd.edu.sg, luwei@sutd.edu.sg`
`{chengchen.xpj, ada.drx, linyan.lll}@alibaba-inc.com`

Abstract

We present the industry dataset information and experimental details of the main paper (Jie et al., 2019) in this supplementary material.

1 Industry Dataset

To justify the robustness of our approach, we also conduct experiments on two datasets from industry: Taobao[1] and Youku[2].

Taobao is an e-commerce site with various types of products. We crawled the product titles from the website and annotate the titles with 9 entity types. Table 1 shows the detailed entity information of the Taobao dataset. The leftmost column represent the categorized entity types (i.e., PATTERN, PRODUCT, BRAND and MISC) we used in the experiments.

Youku is a video-streaming website where a number of videos from various domains are presented. We crawled the video titles from the website and again annotate them with 9 entity types. Table 2 shows the detailed entity information of the Youku dataset. Specifically, we group the entities into three types (i.e., FIGURE, PROGRAM and MISC) for our experiments.

As mentioned in the main paper, we further found that there are more entity labels per sentence on these two industry datasets compare to the standard datasets (i.e., CoNLL-2003 and CoNLL-2002). For example, there are 51% of entity labels per sentence (on average) in Taobao dataset whereas there are only 23% of entity labels per centence (on average) in CoNLL-2003 dataset. Because of the high ratio of entity labels in Taobao and Youku datasets, the missing label CRF (M-CRF) can perform a lot less worse compared to the M-CRF on CoNLL datasets.

Grouped Type	Entity Type	#Entity
PATTERN	Model Type	2,173
PRODUCT	Product Description	5,506
	Core Product	21,958
BRAND	Brand Description	331
	Core Brand	3,430
MISC	Location	1,893
	Person	367
	Literature	814
	Product Specification	2,732

Table 1: The entity information for the Taobao dataset.

Grouped Type	Entity Type	#Entity
FIGURE	Figure	4,402
PROGRAM	Variety Show	1,349
	Movie	1,303
	Animation	3,133
	TV Drama	3,087
MISC	Character	446
	Number	1,022
	Location	523
	Song	640

Table 2: The entity information for the Youku dataset.

2 Experimental Details

This section introduces the settings of the baseline models and our approaches. Throughout the experiments, we apply the bidirectional long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997) networks as our neural architecture (Lample et al., 2016) for conditional random fields (CRF) (Lafferty et al., 2001). Specifically, the hidden size of LSTM is set to 100, hidden size of character-level LSTM is set to 50, dropout is set to 0.5. During training, we use the Adam optimizer (Kingma and Ba, 2015) with a batch size of 1 and clipping value of 5. Our model is trained

---

[1] https://www.taobao.com
[2] https://www.youku.com

Algorithm 1 Partial Perceptron

```
Input: Training data: set of $(\mathbf{x}, \mathbf{y}_p) \in \mathcal{D}$
Output: Model parameters $\mathbf{w}$
$\mathbf{w} = \mathbf{0}$
for $i = 1 \dots T$ do    // $T$ iterations
    for $(\mathbf{x}, \mathbf{y}_p) \in \mathcal{D}$ do
        $\mathbf{z} = \arg\max_{\mathbf{z}} \mathbf{w} \cdot \mathbf{f}(\mathbf{x}, \mathbf{z})$
        if $\mathbf{z} \neq \mathbf{y}_p$ then
            //check positions with gold labels
            update($\mathbf{w}$)
        end if
    end for
end for
```

with 100 epochs. We select the above hyperparameters based on the best performance on development set.

2.1 Baseline Systems

Partial Perceptron We augment the partial perceptron model (Carlson et al., 2009) with BiLSTM as the neural architecture. In this model, we only consider the scores involved the tokens with available (i.e., annotated) labels during the training process. Algorithm 1 shows the procedure of training a partial perceptron.

Transductive Perceptron This model (Fernandes and Brefeld, 2011) augment an additional Hamming loss function during the update procedure in the structured perceptron. Essentially, they applied the max-margin training strategy in the structured perceptron. First, we obtain a pseudo ground-truth label sequence through a constrained Viterbi decoding[3]:

$$\mathbf{y}_{pseudo} = \arg\max_{\mathbf{y} \in \mathcal{C}(\mathbf{y}_p)} \mathbf{w}^{\mathbf{T}} \mathbf{f}(\mathbf{x}, \mathbf{y}) \quad (1)$$

In other words, we first use the current model parameters to obtain a label sequence as pseudo gold sequence for structured perceptron training. Secondly, they obtain the prediction by max-margin decoding procedure:

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y}} [loss(\mathbf{y}_{pseudo}, \mathbf{y}) + \mathbf{w}^{\mathbf{T}} \mathbf{f}(\mathbf{x}, \mathbf{y})] \quad (2)$$

where the loss function is a Hamming loss. Lastly, we perform a perceptron update:

$$\mathbf{w}' = \mathbf{w} + \mathbf{f}(\mathbf{x}, \mathbf{y}_{pseudo}) - \mathbf{f}(\mathbf{x}, \hat{\mathbf{y}}) \quad (3)$$

The Hamming loss in the transductive perceptron is defined as follows:

$$loss(\mathbf{y}_{pseudo}, \mathbf{y}) = \sum_{t=1}^{|\mathbf{y}|} \lambda(t) \quad (4)$$

where $\lambda(t) = \lambda_L$ when t is the time step that involves available labels, and $\lambda(t) = \lambda_U$ when t is the time step that involves unavailable labels. During our experiments, we set $\lambda_L = 1$ and $\lambda_U = 0.1$.

2.2 Implementation Details

We perform iterative training for our *hard* and *soft* approaches. As mentioned in the main paper, we perform iterative training in a k-fold cross-validation manner. Empirically, we set the maximum iteration number to 10, which is enough for our approaches to converge and have a stable q distribution.

References

Andrew Carlson, Scott Gaffney, and Flavian Vasile. 2009. Learning a named entity tagger from gazetteers with the partial perceptron. In *Proceedings of AAAI Spring Symposium: Learning by Reading and Learning to Read.*

Eraldo R Fernandes and Ulf Brefeld. 2011. Learning from partially annotated sequences. In *Proceedings of ECML-KDD.*

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Zhanming Jie, Pengjun Xie, Wei Lu, Ruixue Ding, and Linlin Li. 2019. Better modeling of incomplete annotations for named entity recognition. In *Proceedings of NAACL.*

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of ICLR.*

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of ICML.*

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of NAACL.*

---

[3] This process guarantees the pseudo ground-truth sequence always contains the available labels.