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

Supervised approaches to named entity recognition (NER) are largely developed based on the assumption that the training data is fully annotated with named entity information. However, in practice, annotated data can often be imperfect with one typical issue being the training data may contain incomplete annotations. We highlight several pitfalls associated with learning under such a setup in the context of NER and identify limitations associated with existing approaches, proposing a novel yet easy-to-implement approach for recognizing named entities with incomplete data annotations. We demonstrate the effectiveness of our approach through extensive experiments.[1]

1 Introduction

Named entity recognition (NER) (Tjong Kim Sang, 2002; Tjong Kim Sang and De Meulder, 2003) as one of the most fundamental tasks within natural language processing (NLP) has received significant attention. Most existing approaches to NER focused on a supervised setup, where fully annotated named entity information is assumed to be available during the training phase. However, in practice, obtaining high-quality annotations can be a very laborious and expensive process (Snow et al., 2008). One of the common issues with data annotations is there may be incomplete annotations.

Figure 1 shows an example sentence with two named entities "*John Lloyd Jones*" and "*BBC radio*" of type PER (person) and ORG (organization), respectively. Following the standard BIOES tagging scheme (Ramshaw and Marcus, 1999; Ratinov and Roth, 2009), the corresponding gold label sequence is shown below the sentence. When the data annotations are incomplete, certain labels

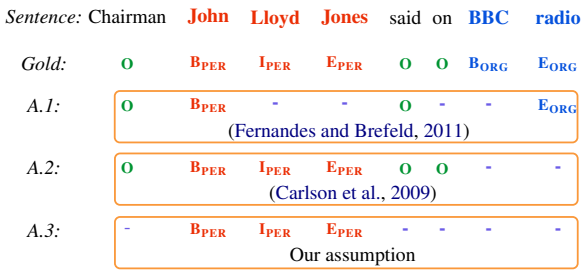


Figure 1: An example sentence with gold named entity annotations and different assumptions (i.e., *A.1* to *A.3*) on *available labels*. "-" represents a missing label.

may be missing from the label sequence. Properly defining the task is important, and we argue there are two possible potential pitfalls associated with modeling incomplete annotations, especially for the NER task.

Several previous approaches assume the incomplete annotations can be obtained by simply removing either word-level labels (Fernandes and Brefeld, 2011) or span-level labels (Carlson et al., 2009). As shown in Figure 1, under both assumptions (i.e., *A.1* and *A.2*), there will be words annotated with O labels. The former approach may even lead to sub-entity level annotations (e.g., "*radio*" is annotated as part of an entity). However, we argue such assumptions can be largely unrealistic. In practice, annotators are typically instructed to annotate named entities for complete word spans only (Settles et al., 2008; Surdeanu et al., 2010). Thus, sub-entity level annotations or O labels [2] should not be assumed to be avail-

[1]Our code and data are available at http://statnlp.org/research/ie.

[2]Why should the O labels be assumed unavailable? This is because the annotators typically do not actively specify the O labels when working on annotations. If the annotator chooses not to annotate a word, it could either mean it is not part of any entity, *or* the word is actually part of an entity but the annotator neglected it in the annotation process (therefore we have incomplete annotations). However, we note that assigning the O label to a word would precisely indicate it is strictly *not* part of any entity, which is not desirable.

Proceedings of NAACL-HLT 2019, pages 729–734
Minneapolis, Minnesota, June 2 - June 7, 2019.

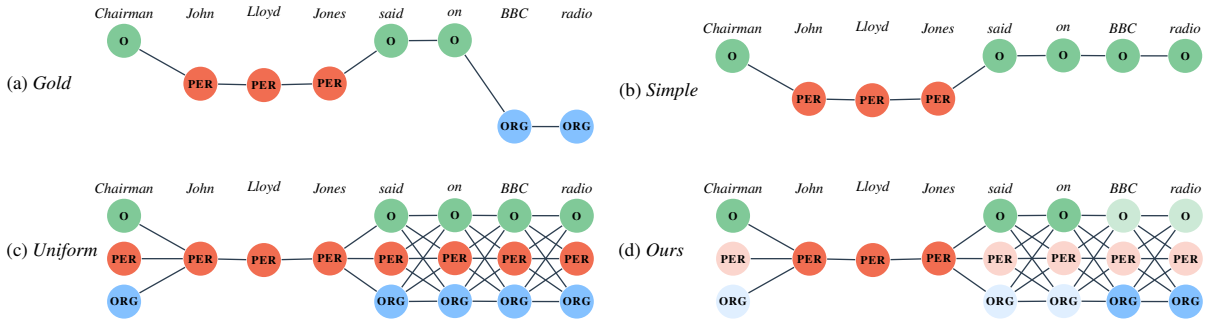


Figure 3: Graphical illustrations on different assumptions on *unavailable labels*, where the entity "*John Lloyd Jones*" of type PER is labeled but "*BBC radio*" of type ORG is missing. Each path refers to one possible complete label sequence, and the density of the color indicates probability (we excluded B and E tags for brevity).

able (*A.3*). Therefore such approaches are making sub-optimal assumptions on the *available labels*.

When the proper assumptions on the available labels are made, one can typically model the missing labels as latent variables and train a latent-variable conditional random fields model (Quattoni et al., 2005). One such approach is presented in (Bellare and McCallum, 2007). Their work focused on the citation parsing[3] (i.e., sequence labeling) task which does not suffer from the above issue as no O label is involved. However, though the approach was shown effective in the citation parsing task, we found its effectiveness does not transfer to the NER task even in the absence of the above available labels issue. As we would highlight later, the reason is related to the undesirable assumptions on the *unavailable labels*.

In this work, we tackle the incomplete annotation problem when building an NER system, under a more realistic yet more challenging scenario. We present a novel, effective, yet easy-to-implement approach, and conduct extensive experiments on various datasets and show our approach significantly outperforms several previous approaches.

2 Related Work

Previous research efforts on partially annotated data are mostly based on the conditional random fields (CRF) (Lafferty et al., 2001), structured perceptron (Collins, 2002) and max-margin (Tsochantaridis et al., 2005) (e.g. structural support vector machine) models. Bellare and McCallum (2007) proposed a missing label linear-chain CRF[4] which is essentially a latent-variable CRF (Quattoni et al., 2005) on citation parsing (McCallum et al., 2000). This model had also been used in part-of-speeching tagging and segmentation task with incomplete annotations (Tsuboi et al., 2008; Liu et al., 2014; Yang and Vozila, 2014). Yang et al. (2018) showed the effectiveness of such a model on Chinese NER with incomplete annotations due to the fact that they required a certain number of fully annotated data to perform joint training. Greenberg et al. (2018) applied this model on a biomedical NER task and achieved promising performance with incomplete annotations. However, in their assumption for the incomplete annotations, the O labels are still considered, which we believe is not realistic. Carlson et al. (2009) modified the structured perceptron algorithm and defined features only on the tokens with annotated labels in partially labeled sequences. Fernandes and Brefeld (2011) and Lou et al. (2012) proposed to use a large-margin learning framework similar to structured support vector machines with latent variables (Yu and Joachims, 2009).

3 Approach

Given the input word sequence $\mathbf{x}$, the NER task is to predict a label sequence $\mathbf{y}$ that encodes the NER information (e.g., in a form following the BIOES tagging scheme). Given a training set that consists of completely labeled data $\mathcal{D}$, one can tackle this problem using a standard linear-chain conditional random field (CRF) (Lafferty et al., 2001) whose loss function is as follows:[5]

$$\mathcal{L}(\mathbf{w}) = -\sum_i \log p_{\mathbf{w}}(\mathbf{y}^{(i)}|\mathbf{x}^{(i)}) \qquad (1)$$

[3]The task is to tag the BibTex records with different labels (i.e., "title", "author", "affiliation" and so on).

[4]This model was also named as Partial CRF (Carlson et al., 2009) and EM Marginal CRF (Greenberg et al., 2018).

[5]In practice, we also have an L_2 regularization term, which we exclude from the formula for brevity.

where $(\mathbf{x}^{(i)}, \mathbf{y}^{(i)})$ is the i-th instance from $\mathcal{D}$.

Now, assume we have an incomplete label sequence $\mathbf{y}_p^{(i)}$. From such a $\mathbf{y}_p^{(i)}$ we should be able to derive a set of all possible complete label sequences that are compatible with (i.e., contain) the incomplete label sequence, and let us call this set $\mathcal{C}(\mathbf{y}_p^{(i)})$. We can rewrite the above function as:

$$\mathcal{L}(\mathbf{w}) = -\sum_i \log \sum_{\mathbf{y} \in \mathcal{C}(\mathbf{y}_p^{(i)})} q_{\mathcal{D}}(\mathbf{y}|\mathbf{x}^{(i)}) p_{\mathbf{w}}(\mathbf{y}|\mathbf{x}^{(i)})$$

We illustrate in Figure 3 several previous approaches as well as our approach. In this example, the entity *BBC radio* of type ORG is not annotated. Figure 3(a) shows a single path that corresponds to the gold label sequence. Figure 3(b) illustrates a naive approach, where we regard all the missing labels as O labels. This essentially assumes that the q distribution in the above equation puts all probability mass to this single label sequence, which is an incorrect assumption.

Now let us look at what assumptions on q have been made by the existing approach of Bellare and McCallum (2007). The model regards the missing labels as latent variables and learns a latent variable CRF using the following loss:

$$-\sum_i \log \sum_{\mathbf{y} \in \mathcal{C}(\mathbf{y}_p^{(i)})} p_{\mathbf{w}}(\mathbf{y}|\mathbf{x}^{(i)}) \qquad (2)$$

The resulting model is called *missing label linear-chain CRF* (*M-CRF*) [6]. As we can see from the above function, this is essentially equivalent to say q is a uniform distribution that assigns equal probabilities to all possible complete label sequences in $\mathcal{C}(\mathbf{y}_p^{(i)})$.

We believe such an assumption on q that describes *unavailable labels* can be improved. As we can see from the above example in Figure 3(d), a more desirable assumption about q is to put more probability mass to a path that is close to the gold path. In practice, their approach worked for the task of citation parsing, where the q distribution may not deviate much from the uniform distribution (Figure 3(c)) in such a task. However, in the task of NER, we find such a simple treatment to the q distribution often leads to sub-optimal results (as we can see in the experiments later) as the q distribution is highly skewed due to the large amount of O labels. This observation motivates us to find a proper way to define q that can approximate the gold label distribution in this work.

[6] Similar assumptions have also been made by (Carlson et al., 2009; Fernandes and Brefeld, 2011), but they used structured perceptron (Collins, 2002) instead.

Dataset	Training		Validation		Test		Entities	
	#entity	#sent	#entity	#sent	#entity	#sent	#	c (%)
CoNLL-2003	23,499	14,041	5,942	3,250	5,648	3,453	4	23.0
CoNLL-2002	18,796	8,322	4,338	1,914	3,559	1,516	4	12.4
Taobao	29,397	6,000	4,941	998	4,866	1,000	4	51.0
Youku	12,754	8,001	1,580	1,000	1,570	1,001	3	41.7

Table 1: Data statistics for the datasets.

3.1 Estimating q

Inspired by the *classifier stacking* technique used in Nivre and McDonald (2008), we empirically found that a reasonable q distribution can be acquired in a k-fold cross-validation fashion.

We first start with an initialization step where we assign specific labels to words without labels, forming complete label sequences (we will discuss our initialization strategy in experiments). Next, we perform k-fold cross-validation on the training set. Specifically, each time we train a model with (k-1) folds of the data and based on the learned model we define our q distribution.

We describe two different ways of defining the q distribution, namely the *hard* approach, and the *soft* approach. In the *hard* approach, the resulting q distribution is a collapsed distribution that assigns probability 1 to a single complete label sequence, whereas in the *soft* approach each possible label sequence will get a certain probability score.

In the *hard* approach, after training a model from (k-1) folds, we apply a constrained Viterbi procedure [7] to the sentences in the remaining fold. In the *soft* approach, we use a constrained version of the forward-backward procedure and calculate the marginal probabilities associated with each label at each unlabeled position. The score of each complete label sequence can then be calculated as a product of all such marginal probabilities. We note that in the above procedure the estimation to q depends on the initialization. Thus we iterate the above procedure, which allows us to converge to an improved q.

4 Experiments

We conduct experiments on two standard NER datasets – CoNLL-2003 English and CoNLL-2002 Spanish datasets that consist of news articles. We

[7] The algorithm will ensure the resulting complete label sequence is compatible with the incomplete label sequence.

Approach	CoNLL-2003			CoNLL-2002			Taobao			Youku		
	P.	*R.*	*F.*	*P.*	*R.*	*F.*	*P.*	*R.*	*F.*	*P.*	*R.*	*F.*
Simple	93.6	68.6	79.2	86.8	57.0	68.8	83.1	46.7	59.8	91.1	49.1	63.8
LSTM-M-CRF	13.0	90.5	22.8	6.2	84.5	11.6	33.0	83.0	47.2	17.6	83.7	29.1
LSTM-Partial Perceptron	27.6	82.0	41.3	22.3	66.3	33.3	26.6	59.7	36.8	19.7	69.0	30.6
LSTM-Transductive Perceptron	11.7	90.5	20.7	6.1	84.3	11.4	32.9	82.2	47.0	16.0	81.9	26.7
Ours (*hard*)	88.1	89.9	89.0	80.8	82.1	81.5	69.3	77.4	73.1	77.2	79.8	78.5
Ours (*soft*)	89.0	90.1	**89.5**	81.3	82.7	**82.0**	69.7	78.1	**73.7**	78.1	79.6	**78.8**
Complete	91.0	90.8	90.9	85.7	85.8	85.8	82.3	82.6	82.4	83.0	81.7	82.4

Table 2: Performance comparison between different baseline models and our approaches on 4 datasets with $\rho = 0.5$ (for *Complete* model, $\rho = 1.0$).

notice that incomplete annotation issue is very common in the industry setup. Therefore we also consider two new datasets from industry – Taobao and Youku datasets[8] consisting of product and video titles in Chinese. We crawled and manually annotated such data with named entities[9]. Table 1 shows the statistics of the datasets. The last two columns show the number of entity types and the percentage of words (i.e., c in Table 1) that are parts of an NE. Based on our assumption on the *available labels* in Section 1, we randomly remove a certain number of entities as well as all O labels and use ρ to represent the ratio of annotated entities. For example, $\rho = 0.6$ means we keep 60% of all the entities and remove the annotations of 40% of the entities. Meanwhile, the O labels are considered unavailable.

We follow Lample et al. (2016) and apply the bidirectional long short-term memory (Hochreiter and Schmidhuber, 1997) (BiLSTM) networks as the neural architecture for all baselines and our approaches. Specifically, we implement the following baselines: a *Simple* model which is a linear-chain LSTM-CRF model and we treat all missing labels as O; the missing label CRF (Bellare and McCallum, 2007; Greenberg et al., 2018) (LSTM-M-CRF) model; the *partial perceptron* (Carlson et al., 2009) model, which is a structured perceptron (Collins, 2002) but only considers the scores on the words with available labels; the *transductive perceptron* (Fernandes and Brefeld, 2011) model where they introduce a Hamming loss function during the perceptron training process; lastly, we train an LSTM-CRF (Lample et al., 2016) with complete annotations as the upper bound (*Complete*). For English and Spanish, we use exactly the same embeddings used in Lample et al. (2016). We train our Chinese character embeddings on the Chinese Gigaword[10] corpus. The resulting implementation achieves 90.9 and 85.8 F-scores on CoNLL-2003 English and CoNLL-2002 Spanish datasets, respectively. These benchmark results are comparable with the results reported in the state-of-the-art NER systems (Lample et al., 2016; Ma and Hovy, 2016; Reimers and Gurevych, 2017).

For initialization in our approaches, we run the *Simple* model on each fold and use the results to initialize our q distribution[11]. Detailed descriptions on experiment settings (e.g., hidden dimension of LSTM and optimizer) and baseline systems are provided in supplementary material.

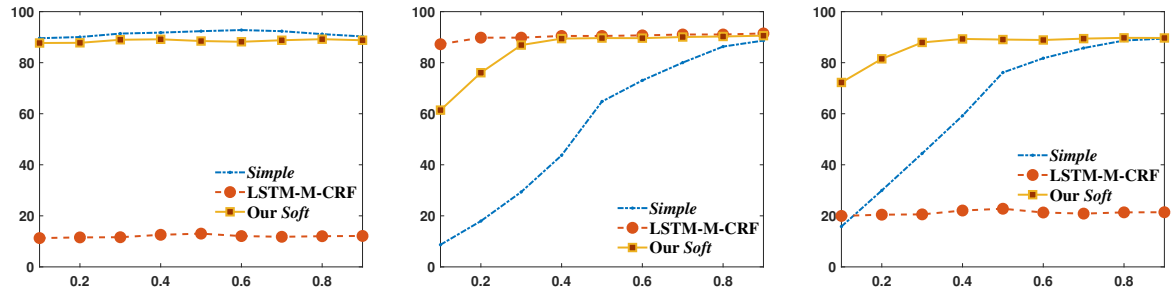
Main Results Table 2 presents the comparisons among all approaches on four datasets with $\rho = 0.5$ and $k = 2$. Our preliminary experiments show that a larger k value have a negligible effect on the results. A similar finding was also reported in Nivre and McDonald (2008). The *Simple* model has high precision and low recall as it treats unknown labels as O. Previous models for incomplete annotations achieve a much lower F-score compared to the *Simple* model and our approaches. Due to their uniform assumption on q over the missing labels, these models typically can recall more entities. The partial perceptron (Carlson et al., 2009) among these three models yields a relatively lower recall as features are not defined over the words with missing labels.

[8] http://www.taobao.com/ and http://www.youku.com/

[9] Details of all datasets can be found in the supplementary material.

[10] https://catalog.ldc.upenn.edu/LDC2003T09

[11] Similar to the EM procedure, a good initialization is crucial for our approach. We found using random initialization can lead to substantially worse results and a better initialization can be used to further improve the results.



(a) Precision (b) Recall (c) F-score

Figure 4: Precision, Recall and F-score with different ρ on CoNLL-2003 dataset.

The difference in F-score between these three models and the *Simple* model is more significant on the two CoNLL datasets than on Taobao and Youku. As shown in Table 1, the latter two datasets have more words labeled as parts of entities (i.e., a higher c). This means these industrial datasets have less O labels, making such baseline models suffer less from their assumptions on the *unavailable labels*. With a properly learned q distribution, our approaches improves the recall score over the *Simple* model while preserving a high precision. Our *soft* approach consistently achieves a better F-score compared with the *hard* approach on all datasets with $p < 0.001$. Compared to the *Complete* upper bound, our *soft* approach are still more than 3% lower in F-score on the CoNLL-2002, Taobao and Youku datasets. However, we can see that the *soft* approach achieves much higher performance compared to this variant on other datasets. We attribute this phenomenon to our approaches' ability in retrieving most of the entities in the training set. Empirically, we found our *soft* approach can recover 94% of the entities in the training set of the CoNLL-2003 dataset.

The overall results show the underlying scenario is challenging for commonly adopted models in handling incomplete annotations and our approaches can achieve better performance compared with them.

Effect of ρ We conduct experiments with different ρ from 0.1 to 0.9 for our *soft* approach against the *Simple* and LSTM-M-CRF models. Figure 4 shows how the precision, recall and F-score on CoNLL-2003 change as we increase ρ. The F-score of the *Simple* baseline increases progressively as ρ increases. LSTM-M-CRF always maintains a low F-score which is not sensitive to different ρ values because of their high recall and low precision values as we can see in Figure 4 (a, b). The improvement of our approach attributes to the increase of recall as the precision is constantly high and stable. We can see that our *soft* approach performs particularly well when ρ is larger than 0.3 which indicates a modest amount of missing labels in practice.

5 Conclusions and Future Work

In this work, we identified several limitations associated with previous assumptions when performing sequence labeling with incomplete annotations, and focused on the named entity recognition task. We presented a novel and easy-to-implement solution that works under a realistic and challenging assumption on the incomplete annotations. Through extensive experiments and analysis, we demonstrated the effectiveness of our approach.

Although we focused on the task of named entity recognition in this work, we believe the proposed approach may find applications in some other sequence labeling tasks or other more general structured prediction problems where the issue of incomplete annotations is involved. We leave them as future work.

Acknowledgments

We would like to thank the anonymous reviewers for their constructive comments on this work. This work is done under a collaborative agreement between SUTD and Alibaba on an Alibaba Innovative Research (AIR) Program funded by Alibaba, where Alibaba provided data and helped with experiments. We appreciate Alibaba's generosity in the agreement that makes it possible for us to make all data and code in this research publicly available upon acceptance of this paper. This work is also partially supported by SUTD project PIE-SGP-AI-2018-01.

References

Kedar Bellare and Andrew McCallum. 2007. Learning extractors from unlabeled text using relevant databases. In *Proceedings of International Workshop on Information Integration on the Web*.

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*.

Michael Collins. 2002. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *Proceedings of EMNLP*.

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

Nathan Greenberg, Trapit Bansal, Patrick Verga, and Andrew McCallum. 2018. Marginal likelihood training of bilstm-crf for biomedical named entity recognition from disjoint label sets. In *Proceedings of EMNLP*.

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

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*.

Yijia Liu, Yue Zhang, Wanxiang Che, Ting Liu, and Fan Wu. 2014. Domain adaptation for crf-based chinese word segmentation using free annotations. In *Proceedings of EMNLP*.

Xinghua Lou, Fred A Hamprecht, and IWR HCI. 2012. Structured learning from partial annotations. In *Proceedings of ICML*.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. In *Proceedings of ACL*.

Andrew Kachites McCallum, Kamal Nigam, Jason Rennie, and Kristie Seymore. 2000. Automating the construction of internet portals with machine learning. *Information Retrieval*, 3(2):127–163.

Joakim Nivre and Ryan McDonald. 2008. Integrating graph-based and transition-based dependency parsers. In *Proceedings of ACL*.

Ariadna Quattoni, Michael Collins, and Trevor Darrell. 2005. Conditional random fields for object recognition. In *Proceedings of NIPS*.

Lance A Ramshaw and Mitchell P Marcus. 1999. Text chunking using transformation-based learning. In *Proceedings of Third Workshop on Very Large Corpora*.

Lev Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of CoNLL*.

Nils Reimers and Iryna Gurevych. 2017. Reporting score distributions makes a difference: Performance study of lstm-networks for sequence tagging. In *Proceedings of EMNLP*.

Burr Settles, Mark Craven, and Lewis Friedland. 2008. Active learning with real annotation costs. In *Proceedings of NIPS Workshop on Cost-Sensitive Learning*.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y Ng. 2008. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of EMNLP*.

Mihai Surdeanu, Ramesh Nallapati, and Christopher Manning. 2010. Legal claim identification: Information extraction with hierarchically labeled data. In *Proceedings of LREC*.

Erik F Tjong Kim Sang. 2002. Introduction to the conll-2002 shared task: Language-independent named entity recognition. In *Proceedings of CoNLL*.

Erik F Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of CoNLL*.

Ioannis Tsochantaridis, Thorsten Joachims, Thomas Hofmann, and Yasemin Altun. 2005. Large margin methods for structured and interdependent output variables. *Journal of machine learning research*, 6:1453–1484.

Yuta Tsuboi, Hisashi Kashima, Shinsuke Mori, Hiroki Oda, and Yuji Matsumoto. 2008. Training conditional random fields using incomplete annotations. In *Proceedings of COLING*.

Fan Yang and Paul Vozila. 2014. Semi-supervised chinese word segmentation using partial-label learning with conditional random fields. In *Proceedings of EMNLP*.

Yaosheng Yang, Wenliang Chen, Zhenghua Li, Zhengqiu He, and Min Zhang. 2018. Distantly supervised ner with partial annotation learning and reinforcement learning. In *Proceedings of COLING*.

Chun-Nam John Yu and Thorsten Joachims. 2009. Learning structural svms with latent variables. In *Proceedings of ICML*.