Towards Improving Neural Named Entity Recognition with Gazetteers

Tianyu Liu[*]
Peking University
ty-liu@pku.edu.cn

Jin-Ge Yao **Chin-Yew Lin**
Microsoft Research Asia
{jinge.yao,cyl}@microsoft.com

Abstract

Most of the recently proposed neural models for named entity recognition have been purely data-driven, with a strong emphasis on getting rid of the efforts for collecting external resources or designing hand-crafted features. This could increase the chance of overfitting since the models cannot access any supervision signal beyond the small amount of annotated data, limiting their power to generalize beyond the annotated entities. In this work, we show that properly utilizing external gazetteers could benefit segmental neural NER models. We add a simple module on the recently proposed hybrid semi-Markov CRF architecture and observe some promising results.

1 Introduction

In the past few years, neural models have become dominant in research on named entity recognition (NER) (Lample et al., 2016; Ma and Hovy, 2016; Chiu and Nichols, 2016, *inter alia*), as they effectively utilize distributed representations learned from large-scale unlabeled texts (Pennington et al., 2014; Peters et al., 2018; Devlin et al., 2018, *inter alia*), while avoiding the huge efforts required for designing hand-crafted features or gathering external lexicons. Results from modern neural NER models have achieved new state-of-the-art performance over standard benchmarks such as the popular CoNLL 2003 shared task dataset (Tjong Kim Sang and De Meulder, 2003).

An end-to-end model with the property of *letting the data speak for itself* seems to be appealing at first sight. However, given that the amount of labeled training data for NER is relatively small when compared with other tasks with millions of training examples, the annotated entities could only achieve a rather limited coverage for a theoretically infinite space of variant entity names. Moreover, current neural architectures heavily rely on the word form due to the use of word embeddings and character embeddings, which could lead to a high chance of overfitting. [1] For instance, all the appearances of the single token *Clinton* in the CoNLL 2003 dataset are person names, while in practice it is also possible to refer to locations.[2] Data-driven end-to-end models trained on that dataset could implicitly bias towards predicting `PERSON` for most occurrences of *Clinton* even under some contexts when it refers to a location.

On the other hand, for frequently studied languages such as English, people have already collected dictionaries or lexicons consisting of long lists of entity names, known as *gazetteers*. Gazetteers could be treated as an external source of knowledge that could guide models towards wider coverage beyond the annotated entities in NER datasets. In traditional log-linear named entity taggers (Ratinov and Roth, 2009; Luo et al., 2015), gazetteers are commonly used as discrete features in the form of whether the current token or current span is appearing in the gazetter or not. There does not seem to be any reason for a neural model not to utilize the off-the-shelf gazetters.

In this paper, we make a simple attempt in utilizing gazetteers in neural NER. Building on a recently proposed architecture called hybrid semi-Markov conditional random fields (HSCRFs) where span-level scores are derived from token-label scores, we introduce a simple additional module that scores a candidate entity span by the degree it softly matches the gazetteer. Experimental studies over CoNLL 2003 and OntoNotes show the utility of gazetteers for neural NER models.

[*] Work during internship at Microsoft Research Asia

[1] In fact, traditional feature-based models also suffer from similar overfitting issues when trained on limited data, but in practice they could be easily spotted and fixed due to the transparency of linear feature weights.

[2] See e.g., `https://en.wikipedia.org/wiki/Clinton_(disambiguation)`

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 5301–5307
Florence, Italy, July 28 - August 2, 2019. ©2019 Association for Computational Linguistics

2 Framework

2.1 Hybrid semi-Markov CRFs

Our approach is by nature based on the hybrid semi-Markov conditional random fields (HSCRFs) proposed by Ye and Ling (2018), which connect traditional CRFs (Lafferty et al., 2001) and semi-Markov CRFs (Sarawagi and Cohen, 2005) by simultaneously leveraging token-level and segment-level scoring information.

Let $\mathbf{s} = \langle s_1, \ldots, s_p \rangle$ denote a *segmentation* of input sequence $\mathbf{x} = \langle x_1, \ldots, x_n \rangle$, where a *segment* $s_j = \langle t_j, u_j, y_j \rangle$ represents a span with a *start position* t_j, an *end position* u_j, and a *label* $y_j \in Y$. We assume that all segments have positive lengths and the start position of the first segment is always 1, then the segmentation $\mathbf{s}$ satisfies $t_1 = 1$, $u_p = n$, $u_j - t_j \geq 0$, and $t_{j+1} = u_j + 1$ for $1 \leq j < p$. Let $\mathbf{l} = \langle l_1, \ldots, l_n \rangle$ be the corresponding token-level labels of $\mathbf{x}$. A traditional semi-CRF (Sarawagi and Cohen, 2005) gives a segmentation of an input sequence and assign labels to each segment in it. For named entity recognition tasks, a correct segmentation of the sentence *Scottish Labour Party narrowly backs referendum* should be $\mathbf{s} = \langle (1, 3, ORG), (4, 4, O), (5, 5, O), (6, 6, O) \rangle$, and the token-level label sequence under a BILOU tagging scheme [3] should become $\mathbf{l} = \langle B{-}ORG, I{-}ORG, L{-}ORG, O, O, O \rangle$.

HSCRFs inherit the definition of segmentation probability from traditional semi-CRFs. Given a sequence $\mathbf{x} = \langle x_1, \ldots, x_n \rangle$, the probability of segmentation $\mathbf{s} = \langle s_1, \ldots, s_p \rangle$ is defined as

$$\Pr(\mathbf{s} \mid \mathbf{x}) = \frac{\text{score}(\mathbf{s}, \mathbf{x})}{Z(x)}, \tag{1}$$

where $\text{score}(\mathbf{s}, \mathbf{x}) = \prod_{j=1}^{p} \psi(y_j, y_{j+1}, \mathbf{x}, t_j, u_j)$, and $Z(x) = \sum_{\mathbf{s}'} \text{score}(\mathbf{s}', \mathbf{x})$ is the normalization term. Note that y_{p+1} is defined as a special $\langle \text{END} \rangle$. The Viterbi algorithm could be used for decoding, i.e., getting the most likely segmentation for a query sentence.

HSCRFs employ a specific method to calculate the segment score using token-level labels, with the score potential function $\psi(\cdot)$ defined as $\psi(y_j, y_{j+1}, \mathbf{x}, t_j, u_j) = \exp(\phi_j + b_{y_j, y_{j+1}})$, where

$$\phi_j = \sum_{i=t_j}^{u_j} \varphi_{\text{token}}^{\text{HSCRF}}(l_i, \mathbf{v}'_i) = \sum_{i=t_j}^{u_j} \mathbf{a}_{l_i}^{\mathsf{T}} \mathbf{v}'_i, \tag{2}$$

and $b_{y_j, y_{j+1}}$ is the segment label transition score from y_j to y_{j+1}, $\varphi_{token}(l_i, w_i)$ calculates the score of the i-th token being classified into token-level label l_i , $\mathbf{v}'_i$ is the feature representation vector of the i-th token x_i, and $\mathbf{a}_{l_i}$ is the weight parameter vector for token label l_i. In HSCRFs, $\mathbf{v}'_i$ is the concatenation of (1) BiLSTM encoded representation $\mathbf{v}_i$, (2) $\mathbf{v}_{u_j} - \mathbf{v}_{t_j}$, and (3) $\mathbf{emb}(i - t_j + 1)$, the position embedding in the segment.

2.2 Gazetteer-enhanced sub-tagger

The most naïve attempt could be treating each gazetteer entity as an additional labeled training sentence, but we found consistently decreased performance in our initial experiments, as this would introduce a shift of label distribution given that the amount of gazetteer entity entries are typically large. Therefore, it seems more natural to utilize gazetteers in a separate module rather than naïvely using them as augmented data.

The structure of HSCRFs makes it straightforward to introduce a scoring scheme for candidate spans based on gazetteers. Following the scoring scheme of HSCRFs, we train a span classifier in the form of a sub-tagger and extract token-level features at the same time. Let $\mathbf{z} = \langle z_1, \ldots, z_k \rangle$ be an entity in the gazetteer with a corresponding label $\mathbf{m}$. This span-level label can be expanded into token-level labels $m_1, \ldots, m_k$. For example, the entity *Scottish Labour Party* is labeled as $\langle B{-}ORG, I{-}ORG, L{-}ORG \rangle$ and *Berlin* is labeled as $\langle U{-}LOC \rangle$ under the BILOU scheme. Similar to Equation 2, the scoring function of our sub-tagger is defined as

$$\phi(\mathbf{m}, \mathbf{z}) = \sum_{i=1}^{k} \varphi_{\text{token}}^{\text{subtagger}}(m_i, z_i) = \sum_{i=1}^{k} \mathbf{w}_{m_i}^{\mathsf{T}} \mathbf{v}'_i \tag{3}$$

where $\mathbf{v}'_i$ is defined in Section 2.1 and $\mathbf{w}_{m_i}$ is the weight parameter vector for token label m_i. We calculate $\text{sigmoid}\big(\phi(\mathbf{m}, \mathbf{z})\big)$ as the probability of category $\mathbf{m}$ and minimize the cross-entropy loss for training this sub-tagger.

The token-level BILOU scores derived from the sub-tagger are larger at scale. We rescale the scores with the $\tanh$ activation function

[3] In the BILOU scheme, a model should learn to identify the **B**eginning, the **I**nside and the **L**ast tokens of multi-token chunks as well as **O**utside tokens and **U**nit-length chunks.

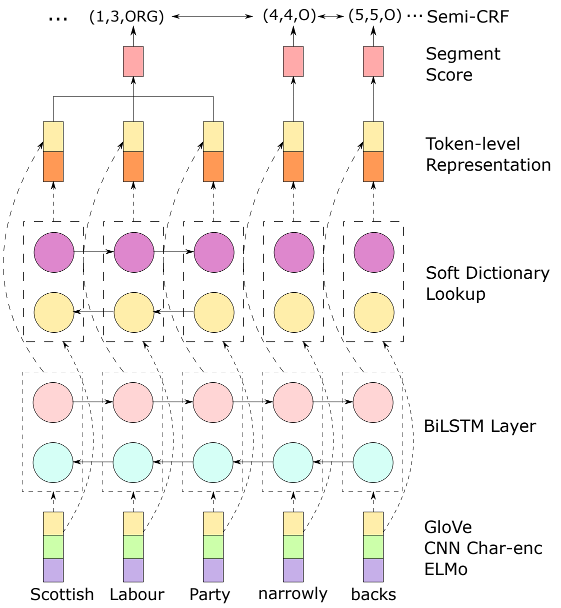

Figure 1: Overall architecture

and concatenate them with the corresponding token representation $\mathbf{v}'_i$ (defined in Section 2.1). Thus, an additional soft dictionary feature vector $\boldsymbol{\eta}_i = \bigoplus_{m \in \mathbf{M}} \tanh\left(\varphi^{\text{subtagger}}_{\text{token}}(m, z_i)\right)$ is derived for each token in a segment, where $\bigoplus$ is the concatenation operation and $\mathbf{M}$ is the set of all BILOU scheme token-level labels. The final ϕ_j for soft dictionary enhanced HSCRF is:

$$\phi_j = \sum_{i=t_j}^{u_j} \varphi^{\text{softdict}}_{\text{token}}(l_i, \boldsymbol{\mu}_i) = \sum_{i=t_j}^{u_j} \mathbf{b}_{l_i}^{\mathsf{T}} \boldsymbol{\mu}_i, \qquad (4)$$

where $\boldsymbol{\mu}_i = \boldsymbol{\eta}_i \bigoplus \mathbf{v}'_i$ and $\mathbf{b}_{l_i}^{\mathsf{T}}$ is the new weight parameter for token label l_i.

The HSCRF model and the sub-tagger derived from it are linear in the way they calculate the span scores. Unlike other semi-CRF models (Zhuo et al., 2016; Zhai et al., 2017; Sato et al., 2017) which utilize neural approaches to derive span scores from word-level representations, HSCRF calculates span score by summing up word-level scores inside a span along *BILOU paths* constrained by tag m_i's.

This sub-tagger could be analogously treated as playing the role of *soft dictionary look-ups*, as opposed to the traditional way that activates a discrete feature only for hard token/span matches.

3 Experiments

3.1 Gazetteers

We use the gazetteers contained in the publicly available UIUC NER system (Khashabi et al., 2018). The gazetteers were originally collected from the web and Wikipedia, consisting of around 1.5 million entities grouped into 79 fine-grained categories. We trimmed and mapped these groups into CoNLL-formatted NER tags (see Appendix for details) with about 1.3 million entities kept.

3.2 Dataset

Evaluation is performed on the CoNLL-2003 English NER shared task dataset (Tjong Kim Sang and De Meulder, 2003) and the OntoNotes 5.0 dataset (Pradhan et al., 2013). We follow the standard train/development/test split described in the original papers along with previous evaluation settings (Chiu and Nichols, 2016).

3.3 Training

Due to the space limit, we leave hyperparameter details to the supplementary materials. [4]

Word representation The representation for a word consists of three parts: pretrained 50-dimensional GloVe word embedding (Pennington et al., 2014), contextualized ELMo embedding (Peters et al., 2018), along with a convolutional character encoder trained from randomly initialized character embeddings, following previous work (Ye and Ling, 2018).

Gazetteer-enhanced sub-tagger We randomly split the gazetteer entities for training (80%) and validation (20%), and sampled 1 million non-entity n-grams (the maximal n is 7) from the CoNLL 2003 training set excluding named entities as negative samples (O labels). We applied early stopping on validation loss when training the sub-tagger.

3.4 Alternative baselines with gazetteers

Many previous NER systems (Ratinov and Roth, 2009; Passos et al., 2014; Chiu and Nichols, 2016) make use of discrete gazetteer features by directly concatenating them with word-level representations. Apart from simple discrete feature concatenation, we also compare our framework with another baseline that utilizes gazetteer embedding as

[4] Our implementation is available at: https://github.com/lyutyuh/acl19_subtagger

an additional feature. We add a single embedding layer for discrete gazetteer features. To be more specific, if a text span corresponds to multiple tags in the gazetteer, we sum all the embedded vector as the final gazetteer tag representation. Otherwise, if a text span has no corresponding tags in the gazetteer, a zero vector of the same dimension will be chosen. Then, the gazetteer tag representation is concatenated with each word-level representation inside a span.

3.5 Results

Table 1 shows the results on the CoNLL 2003 dataset and OntoNotes 5.0 dataset respectively. HSCRFs using gazetteer-enhanced sub-tagger outperform the baselines, achieving comparable results with those of more complex or larger models on CoNLL 2003 and new state-of-the-art results on OntoNotes 5.0. We also attached some out-of-domain analysis in the Appendix.

Model	**Test Set F1-score(±std)**	
	CoNLL	**OntoNotes**
Ma and Hovy (2016)	91.21	-
Lample et al. (2016)	90.94	-
Liu et al. (2018)	91.24±0.12	-
Devlin et al. (2018)	92.8	-
Chiu and Nichols (2016) [5]	91.62±0.33	86.28±0.26
Ghaddar and Langlais ’18	91.73±0.10	87.95±0.13
Peters et al. (2018)	92.22±0.10	89.04±0.27
Clark et al. (2018)	92.6 ±0.1	88.8±0.1
Akbik et al. (2018)	93.09±0.12	89.71
HSCRF	92.54±0.11	89.38±0.11
HSCRF + concat	92.52±0.09	89.73±0.19
HSCRF + gazemb	92.63±0.08	89.77±0.20
HSCRF + softdict	92.75±0.18	89.94±0.16

Table 1: Results on CoNLL 2003 and OntoNotes 5.0

To better attribute the improments of our model, we split the test sets into four non-overlapped subsets according to whether an entity appears in the train set and gazetteer or not, and collect results respectively. We evaluate the performance of our systems on these subsets. Details of the evaluation of each system are shown in Table 2 and Table 3.

We observe that our current approach of sub-tagger soft-dictionary matching consistently improves over baseline approaches on most subsets, while direct concatenating discrete gazetteer features or using gazetteer embedding have sometimes decrease the performance. However, the results on CoNLL and OntoNotes reveal slightly different patterns for the feature concatenation baseline and the gazetteer embedding baseline, making it difficult to analyze the underlying reasons. We leave more systematic experimental studies over the baselines to future work.

We also evaluate the gazetteer sub-tagger on the held-out data of the gazetteer to analyze the potential impact of this module. For predictions, we choose the labels with the highest possibility. If none of the label receives a probability greater than 50%, the sample will be labeled as not being an entity. The results are reported in Table 4.

We can see that while the sub-tagger module could help a lot in identifying person names (`PER`) and organization names (`ORG`), currently the worst-performing category is the miscellaneous type (`MISC`), which is possibly a result of the diversity in this category. Improving the prediction of such entities might further provide performance gains for named entity recognition in general.

4 Discussion

Experimental results demonstrate the usefulness of gazetteer knowledge and show some promising results from our initial attempt to make use of gazetteer information. The sub-tagger has an advantage over hard matching with the capability of recognizing entity names not appearing in but being similar to those contained in the gazetteer. Table 5 lists some examples that the baselines failed to recognize as a complete entity name, while the sub-tagger enhanced system managed to do it. We checked a few cases for which only the sub-tagger enhanced model got correct predictions, and found terms with similar patterns from the gazetteer while not in training data as in Table 6. The gazetteer possesses an abundance of similar terms that enables generalization to out-of-gazetteer items.

In summary, we show that gazetteer-enhanced modules could be useful for neural NER models. Future directions will include trying similarly enhanced modules on other different types of segmental models (Kong et al., 2016; Liu et al., 2016; Zhuo et al., 2016; Zhai et al., 2017; Sato et al., 2017), along with richer representations for further gain. Also, we would like to further explore the possibility to use domain-specific gazetteers or dictionaries to boost the performance of NER in

[5] This work also introduced discrete gazetteer features. We tried their scheme on our gazetteer but we only found consistently decreased performance over the baseline HSCRF.

Model	**Subset** (number of entities with proportions)											
	neither			gazetteer only			training set only			both		
	2042 (36.5%)			655 (11.7%)			1765 (31.5%)			1131 (20.2%)		
HSCRF	84.41	81.37	82.86	97.26	98.38	97.82	96.72	99.09	97.89	96.58	99.84	98.18
HSCRF+gazemb	85.07	81.72	83.36	96.21	98.57	97.38	96.85	99.06	97.94	96.42	99.85	98.11
HSCRF+concat	85.29	81.34	83.27	96.11	98.68	97.38	96.90	99.35	98.11	96.37	99.91	98.11
HSCRF+softdict	84.93	82.16	83.52	97.40	98.53	97.96	97.07	99.31	98.18	96.54	99.91	98.19

Table 2: Detailed test set performance (Precision, Recall, F1) on CoNLL.

Model	**Subset** (number of entities with proportions)											
	neither			gazetteer only			training set only			both		
	2765 (36.5%)			720 (9.5%)			3601 (47.6%)			470 (6.2%)		
HSCRF	80.15	70.42	74.97	95.31	96.48	95.89	92.55	98.91	95.62	95.46	99.66	97.52
HSCRF+gazemb	80.41	71.41	75.64	94.70	96.53	95.60	92.38	98.91	95.53	95.15	99.48	97.27
HSCRF+concat	80.29	72.13	75.99	95.82	96.71	96.26	93.16	98.95	95.97	95.13	99.52	97.27
HSCRF+softdict	80.58	73.36	76.80	96.38	96.46	96.42	93.25	98.96	96.01	95.80	99.62	97.67

Table 3: Detailed test set performance (Precision, Recall, F1) on OntoNotes.

Tag	**Type**		
	Precision	Recall	F1
PER	96.73	97.08	96.91
LOC	83.98	86.20	85.08
ORG	94.99	87.09	90.87
MISC	87.11	72.02	78.85
Overall	94.39	92.65	93.51

Table 4: Sub-tagger evaluation by category. We report the overall recall, precision, and F1 scores of the CoNLL tag set sub-tagger.

HSCRF+softdict	$\underbrace{\text{U.N. Interim Force in Lebanon}}_{\text{ORG}}$
HSCRF+gazemb	$\underbrace{\text{U.N. Interim Force}}_{\text{ORG}}$ in $\underbrace{\text{Lebanon}}_{\text{LOC}}$
HSCRF	$\underbrace{\text{U.N. Interim Force}}_{\text{ORG}}$ in $\underbrace{\text{Lebanon}}_{\text{LOC}}$
HSCRF+softdict	$\underbrace{\text{Hector “Macho” Camacho}}_{\text{PER}}$
HSCRF+gazemb	$\underbrace{\text{Hector “Macho}}_{\text{PER}}$” $\underbrace{\text{Camacho}}_{\text{PER}}$
HSCRF	$\underbrace{\text{Hector}}_{\text{PER}}$ “ $\underbrace{\text{Macho}}_{\text{PER}}$” $\underbrace{\text{Camacho}}_{\text{PER}}$
HSCRF+softdict	$\underbrace{\text{Bodman, Longely \& Dahling}}_{\text{ORG}}$
HSCRF+gazemb	$\underbrace{\text{Bodman}}_{\text{PER}}$, $\underbrace{\text{Longely \& Dahling}}_{\text{ORG}}$
HSCRF	$\underbrace{\text{Bodman}}_{\text{PER}}$, $\underbrace{\text{Longely \& Dahling}}_{\text{ORG}}$

Table 5: Examples from CoNLL 2003 dev set that the soft-dictionary enhanced model classified correctly while other baselines failed.

U.N. Interim Force in Lebanon	Special Security Force Bangladesh Islamic Army in Iraq Grand Army of the Republic
Hector “Macho” Camacho	Charles “Charlie” White Carlos “Carlão” Santos Orlando “Cachaito” López
Bodman, Longely & Dahling	Ransomes, Sims & Jefferies Cravath, Swaine & Moore Drinker, Biddle & Reath

Table 6: Terms similar to CoNLL 2003 dev set entities appearing in the gazetteer.

various domains (Shang et al., 2018), beyond the standard corpora.

Acknowledgement

We thank all the anonymous reviewers for helpful suggestions, especially Reviewer #1 for the thorough comments containing more than 1,500 words in total, from which many points proved to be valuable for improving our initial draft.

References

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649. Association for Computational Linguistics.

Jason Chiu and Eric Nichols. 2016. Named entity recognition with bidirectional lstm-cnns. *Transactions of the Association for Computational Linguistics*, 4:357–370.

Kevin Clark, Minh-Thang Luong, Christopher D. Manning, and Quoc Le. 2018. Semi-supervised sequence modeling with cross-view training. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1914–1925. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Abbas Ghaddar and Phillippe Langlais. 2018. Robust lexical features for improved neural network named-entity recognition. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1896–1907. Association for Computational Linguistics.

Daniel Khashabi, Mark Sammons, Ben Zhou, Tom Redman, Christos Christodoulopoulos, Vivek Srikumar, Nickolas Rizzolo, Lev Ratinov, Guanheng Luo, Quang Do, Chen-Tse Tsai, Subhro Roy, Stephen Mayhew, Zhili Feng, John Wieting, Xiaodong Yu, Yangqiu Song, Shashank Gupta, Shyam Upadhyay, Naveen Arivazhagan, Qiang Ning, Shaoshi Ling, and Dan Roth. 2018. CogCompNLP: Your Swiss Army Knife for NLP. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan.

Lingpeng Kong, Chris Dyer, and Noah A. Smith. 2016. Segmental recurrent neural networks. In *International Conference on Learning Representations*.

John D Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 282–289.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270. Association for Computational Linguistics.

Liyuan Liu, Jingbo Shang, Xiang Ren, Frank Xu, Huan Gui, Jian Peng, and Jiawei Han. 2018. Empower sequence labeling with task-aware neural language model. In *AAAI Conference on Artificial Intelligence*.

Yijia Liu, Wanxiang Che, Jiang Guo, Bing Qin, and Ting Liu. 2016. Exploring segment representations for neural segmentation models. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pages 2880–2886.

Gang Luo, Xiaojiang Huang, Chin-Yew Lin, and Zaiqing Nie. 2015. Joint entity recognition and disambiguation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 879–888, Lisbon, Portugal. Association for Computational Linguistics.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074. Association for Computational Linguistics.

Alexandre Passos, Vineet Kumar, and Andrew McCallum. 2014. Lexicon infused phrase embeddings for named entity resolution. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 78–86, Ann Arbor, Michigan. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237. Association for Computational Linguistics.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. 2013. Towards robust linguistic analysis using ontonotes. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 143–152. Association for Computational Linguistics.

Lev Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, CoNLL '09, pages 147–155, Stroudsburg, PA, USA. Association for Computational Linguistics.

Sunita Sarawagi and William W Cohen. 2005. Semi-markov conditional random fields for information extraction. In *Advances in neural information processing systems*, pages 1185–1192.

Motoki Sato, Hiroyuki Shindo, Ikuya Yamada, and Yuji Matsumoto. 2017. Segment-level neural conditional random fields for named entity recognition. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 97–102, Taipei, Taiwan.

Jingbo Shang, Liyuan Liu, Xiaotao Gu, Xiang Ren, Teng Ren, and Jiawei Han. 2018. Learning named entity tagger using domain-specific dictionary. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2054–2064, Brussels, Belgium. Association for Computational Linguistics.

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 the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*.

Zhixiu Ye and Zhen-Hua Ling. 2018. Hybrid semi-markov crf for neural sequence labeling. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 235–240. Association for Computational Linguistics.

Feifei Zhai, Saloni Potdar, Bing Xiang, and Bowen Zhou. 2017. Neural models for sequence chunking. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Jingwei Zhuo, Yong Cao, Jun Zhu, Bo Zhang, and Zaiqing Nie. 2016. Segment-level sequence modeling using gated recursive semi-Markov conditional random fields. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1413–1423, Berlin, Germany. Association for Computational Linguistics.