

Nested Named Entity Recognition for Chinese Electronic Health Records with QA-based Sequence Labeling

Yu-Lun Chiang¹, Chih-Hao Lin¹, Cheng-Lung Sung¹, and Keh-Yih Su²

¹Data Intelligence R&D Division, CTBC Bank, Co., Ltd, Taipei, Taiwan

²Institute of Information Science, Academia Sinica, Taiwan

{ylchiang914, mr.chihhaolin}@gmail.com

alan.sung@ctbcbank.com, kysu@iis.sinica.edu.tw

Abstract

This study presents a novel QA-based sequence labeling (QASL) approach to naturally tackle both flat and nested Named Entity Recognition (NER) tasks on a Chinese Electronic Health Records (CEHRs) dataset. This proposed QASL approach parallelly asks a corresponding natural language question for each specific named entity type. It then identifies those associated NEs of the same specified type with the BIO tagging scheme. The associated nested NEs are then formed by overlapping the results of various types. Compared with those pure sequence-labeling (SL) approaches, since the given question includes significant prior knowledge about the specified entity type and the capability of extracting NEs with different types, the nested NER task is thus improved, obtaining 90.70% of F1-score. Besides, compared to the pure QA-based approach, our proposed approach retains the SL features, which could extract multiple NEs with the same types without knowing the exact number of NEs in the same passage in advance. Eventually, experiments on our CEHR dataset demonstrate that QASL-based models greatly outperform the SL-based models by 6.12% to 7.14% of F1-score.

Keywords: Nested Named Entity Recognition, Chinese Electronic Health Records, QA-based Sequence Labeling

1 Introduction

Electronic health records (EHRs) contain rich medical information and treatment histories of patients (e.g., various event dates, diagnoses, and treatments). It is beneficial to understand the patients' conditions that all clinicians are



Figure 1: A common example of Chinese electronic health records (CEHRs).

involved in their care. In the past, this information was embedded in unstructured raw texts and extracted manually to databases. Therefore, Named Entity Recognition (NER) task, effectively identifying meaningful named entities (NEs) from unstructured raw texts, has emerged as a hot topic among researchers and practitioners these days.

In Chinese EHRs, a phenomenon often exists that NEs are overlapped or nested, especially in event date types. For example, as shown in Figure 1, The entity ("西元 2019 年 10 月 5 日", Oct. 5, 2019) in the passage has several roles such as the admission date and the emergency date. However, most models only focus on handling flat NER in which NEs do not overlap each other; only a few of them deal with nested NER in which overlapped NEs are allowed.

The NER task has been treated as a sequence labeling (SL) problem in previous works (Lafferty et al., 2001; Hammerton, 2003; Ratnov and Roth, 2009; Collobert et al., 2011; Huang et al., 2015; Ma and Hovy, 2016; Peters et al., 2018; Devlin et al., 2019). With this approach, flat (non-overlapping) NEs within a given passage could be simultaneously identified; however, they failed to detect nested NEs.

To address the issues, various approaches

have been proposed to solve both flat and nested NER with public datasets such as ACE2004 (Doddington et al., 2004), ACE2005 (Christopher Walker et al., 2006), GENIA (Kim et al., 2003), and NNE (Ringland et al., 2019). First, stack-based approaches utilize flat NER layers to sequentially extract entities from inner to outer or outer to inner (Alex et al., 2007; Ju et al., 2018; Wang et al., 2020a). Secondly, graph-based approaches apply constituency parse trees (Finkel and Manning, 2009), hypergraphs (Lu and Roth, 2015; Wang and Lu, 2018; Katiyar and Cardie, 2018), or bipartite graphs (Luo and Zhao, 2020) to identify nested NEs. Thirdly, region-based approaches decompose NER to two stages: detect all possible spans and classify them into pre-defined entity types (Xu et al., 2017; Fisher and Vlachos, 2019; Xia et al., 2019; Zheng et al., 2019; Wang et al., 2020b). Different from public datasets, our Chinese EHR dataset only contains flat NEs and nested NEs with different entity types, meaning that nested NEs with the same types are not in our consideration. Therefore, many above attempts are not the most suitable and intuitive methods for our CEHR dataset due to their complicated models or frameworks.

This study proposed a simple and effective framework of Question Answering Sequence Labeling (QASL). Inspired by Li et al., 2020 (Li et al., 2020), we also re-formalize the NER task to a Question Answering (QA) problem to naturally tackle both flat and nested NER. However, different from this work (Li et al., 2020), we modified the strategy of span selection from predicting start and end positions of entity spans to directly assigning BIO labels to tokens in the input passage. To be more specific, the QASL approach first adopts the corresponding string of the specified NE-type as the query. It then identifies NEs with the BIO tagging scheme by parallelly querying the corresponding NE-type-string (e.g. “入院日期,” Admission Date) for each specific NE type. As shown in Figure 2, the QASL first assigns BIO labels (i.e., Begin (B), Inside (I), or Other (O)) (Ramshaw and Marcus, 1999) to the passage based on a given query/type (“入院日期,” Admission Date). According to the assigned BIO labels, the NE-date (“西元 2019

年 10 月 5 日”, Oct. 5, 2019) is thus identified. Afterward, the QASL conducts the same procedure based on another query/type (“急诊日期,” Emergency Date), and thus identify the same entity with a different type. Last, by conducting the above procedure, all NEs in the passage could be extracted whether they are overlapped or not.

The modification of the span selection strategy has two advantages: (1) BIO labels implicitly tell models the start and end positions of entities and contain rich information among tokens (Wang et al., 2020b) for models. (2) BIO tagging scheme is simple and effective methods to select multiple spans for QA (Segal et al., 2020). It can do well no matter models know how many NEs exist in advance according to questions.

In summary, the contributions of this paper are:

- We propose a novel QA-based sequence labeling (QASL) approach to naturally deal with both flat and nested NER.
- We present the first work to handle the Chinese electronic health records (CEHRs) dataset for both flat and nested NER (To the best of our knowledge).
- We conduct the experiments on a CEHR dataset to show that the proposed QASL is effective.

2 QA-based Sequence Labeling

2.1 Task Formulation

Given a passage $S = \{s_1, s_2, \dots, s_n\}$, where n is the length of the passage, find all the named entities in S with various entity types (according to a pre-specified type-set) $E = \{e_1, e_2, \dots, e_m\}$, where m is the number of entity types. In the framework of QA-based Sequence Labeling (QASL), for each entity type $e \in E$, it is firstly mapped into a predefined query $q_e = \{q_1, q_2, \dots, q_k\}$, where k is the length of query. Then, for each $q_e \in Q$, we find the corresponding named entities (with the same specified type) in S by simply labeling s_i as $l_i \in L = \{B, I, O\}$ according to the BIO scheme (Ramshaw and Marcus, 1999). The associated nested named entities are then formed by overlapping the NER result of each type.

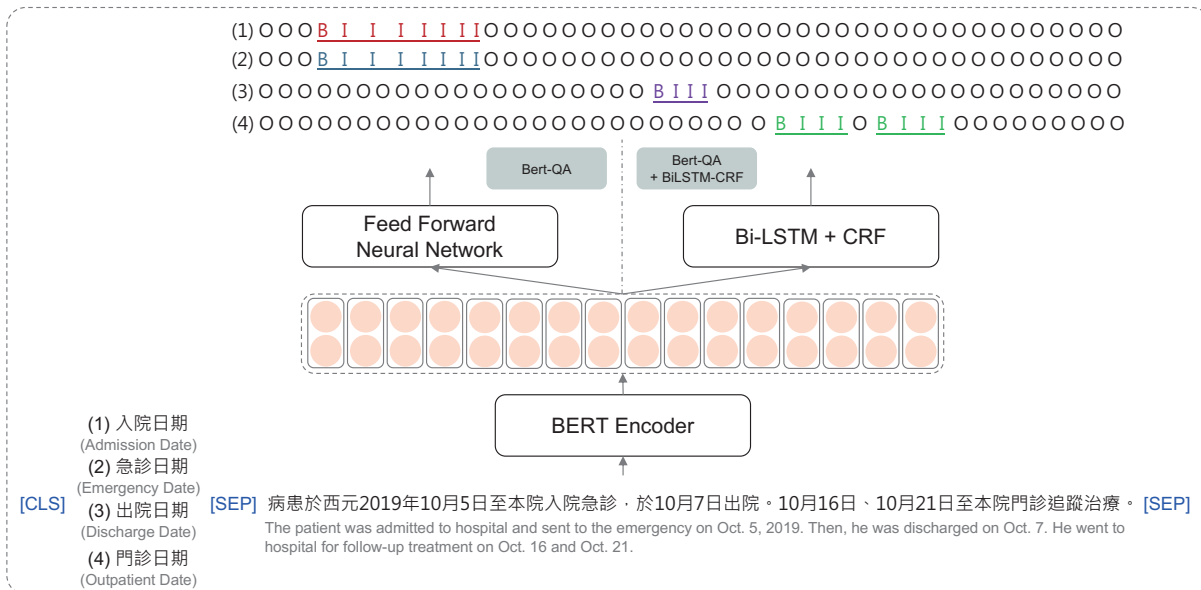


Figure 2: An overview of proposed Question Answering Sequence Labeling (QASL) framework.

2.2 Proposed QASL Model

2.2.1 Query Generation

Since the question could encode prior knowledge about entity types and significantly influence the final results, it is important to generate appropriate questions. To generate the benchmark questions, Li et al. (Li et al., 2020) adopted the Annotation Guideline Notes (e.g., Find locations in the text, including non-geographical locations, mountain ranges, and bodies of water.) to construct the required training data. They achieved the highest F1-score on English OntoNotes 5.0. However, it would not only require an expensive cost to generate the benchmark questions following the guidelines manually, but the questions generated by the guidelines also remain unknown to utilize for another dataset. To avoid those drawbacks mentioned above, we let the questions be keywords (i.e., Chinese NE-Types) in this study, as shown in Table 1. The questions can be easily transformed into the name of entity types, and they can be utilized by different datasets. Therefore, it does not require manual generation, which is expensive, and it is easily generalized by different datasets.

2.2.2 Input Layer

In this paper, we use BERT with whole word masking (BERT-wwm) as the backbone model (Cui et al., 2019). Follow the typical setup

(Li et al., 2020), the question q_e and the passage S are concatenated with the special tokens $[CLS]$ and $[SEP]$, as shown in Figure 2. Then, word embeddings, segmentation embeddings, and positional embeddings for each token are summed together to generate final input representations.

2.2.3 BERT Encoder

The adopted BERT encoder consists of 12 Transformer blocks and 12 self-attention heads by taking the input representation from the input layer and then outputting a context representation. Different from the original BERT (Devlin et al., 2019), BERT-wwm focuses on Chinese language by pre-training with whole word masking (Cui et al., 2019). We only use the passage representations $C \in \mathbb{R}^{n \times d_1}$ from the last hidden layer of BERT-wwm, where d_1 is the dimension with a default value 768 and n is the length of the passage.

2.2.4 Output Layer

This study tests two different structures of output layers: a softmax classifier and a BiLSTM-CRF layer. First, the softmax classifier is that the model predicts the conditional probability distributions P overall categorical labels $L = \{B, I, O\}$, given the passage representations C from BERT encoder:

$$P(L|C; \theta) = \text{softmax}(C \cdot V) \in \mathbb{R}^{n \times 3} \quad (1)$$

Abb.	Entity Type	Abb.	Entity Type	Abb.	Entity Type
ADD	AdmissionDate	OPD	OutpatientDate	RTD	RadiotherapyDate
DCD	DischargeDate	OPDS	OutpatientDateStart	RTDS	RadiotherapyDateStart
ICD	InIntensiveCareDate	OPDE	OutpatientDateEnd	RTDE	RadiotherapyDateEnd
OCD	OutIntensiveCareDate	OPC	OutpatientCount	RTC	RadiotherapyCount
IBD	InBurnWaeDate	EMD	EmergencyDate	CTD	ChemotherapyDate
OBD	OutBurnWaeDate	EMDS	EmergencyDateStart	CTDS	ChemotherapyDateStart
IND	InNegativePressureDate	EMDE	EmergencyDateEnd	CTDE	ChemotherapyDateEnd
OND	OutNegativePressureDate	EMC	EmergencyCount	CTC	ChemotherapyCount
SGN	SurgeryName	SGD	SurgeryDate	SGDE	SurgeryDateEnd
DTN	Drug/TreatmentName	SGDS	SurgeryDateStart	SGC	SurgeryCount
DPN	DepartmentName				

Table 1: The names and abbreviation of entity types.

where θ is the set of all trainable parameters in the model. $V \in \mathbb{R}^{d_1 \times 3}$ is also the trainable parameter. On the other hand, the BiLSTM-CRF first outputs the concatenated hidden representations $H \in \mathbb{R}^{n \times d_2}$ given the passage representations C from BERT encoder, where d_2 is also the dimension with a value of 768. For each $h_i \in H$ and $c_i \in C$:

$$h_i = [\vec{h}_i; \overleftarrow{h}_i] \quad (2)$$

$$\vec{h}_i = LSTM(c_i, \vec{h}_{i-1}; \vec{\theta}) \quad (3)$$

$$\overleftarrow{h}_i = LSTM(c_i, \overleftarrow{h}_{i-1}; \overleftarrow{\theta}) \quad (4)$$

where $\vec{\theta}$, $\overleftarrow{\theta}$ are the trainable parameters in BiLSTM. Besides, the CRF layer (Lafferty et al., 2001) defines the probability of the predicted BIO label sequence Y given the input label sequence X transformed from a given passage S :

$$P(Y|X; \theta) = \frac{e^{\text{score}(X,Y)}}{\sum_{Y'} e^{\text{score}(X,Y')}} \quad (5)$$

The score (Lample et al., 2016) is defined as the sum of transitions and emissions from the BiLSTM:

$$\text{score}(X, Y) = \sum_{i=0}^{n-1} Tr_{y_i, y_{i+1}} + \sum_{i=1}^n Em_{y_i} \quad (6)$$

where Tr is a transition matrix in which $Tr_{y_i, y_{i+1}}$ is the transition parameter from the label y_i to the y_{i+1} . Em is an emission matrix where Em_{y_i} represents the scores of the label y_i at the i -th position.

$$Em = H \cdot U \in \mathbb{R}^{n \times 3} \quad (7)$$

where $U \in \mathbb{R}^{d_2 \times 3}$ is the trainable parameters.

At test time in the structure of the softmax classifier, we take the labels with the largest probability as the predicted results.

$$Y^* = \text{argmax}(P(L|C; \theta)) \in \mathbb{R}^{n \times 1} \quad (8)$$

At test time in the structure of BiLSTM-CRF, we take the label sequence with the largest score as the predicted results by applying the Viterbi algorithm (Viterbi, 1967).

$$Y^* = \text{argmax}(\text{score}(X, Y')) \in \mathbb{R}^{n \times 1} \quad (9)$$

3 Experiments

3.1 Dataset

In this paper, all the experiments are conducted on our Chinese electronic health records (CEHR) dataset.¹ The CEHR dataset is annotated with SQuAD-like style by several well-trained annotators. It is a set of (Passage, Queries, Answers). There are 31 entity types in the CEHR dataset, as shown in Table 1. We extracted that dataset with only flat NEs from the original CEHR dataset as a flat NER dataset, and we took the original CEHR as a nested NER dataset. In the flat NER dataset, the number of passages is 4,328, and the average length of these passages is 70.43. The number of flat NE in these passages is 21,616. On the other hand, in the nested NER dataset, the number of passages is 7,907, and the average length of these passages is 76.08. The number of flat and nested NEs in these passages is 43,577 and 6,978, respectively. Eventually, the flat NER dataset and nested NER dataset are split for training, development, and test set with the ratio 8:1:1.

¹The personal privacy information of all patients in CEHR has been de-identified during the labeling stage.

Model	P	R	F1
Bert	95.45	96.33	95.89
-BiLSTM-CRF	95.37	96.46	95.91
Bert-QA	94.24	95.23	94.73
-BiLSTM-CRF	95.06	95.98	95.52

Table 2: Model Performance on flat NER.

3.2 Baselines and Parameter Settings

In this study, we propose and test two different kinds of QASL-based models: BERT-QA and BERT-QA-BiLSTM-CRF. For comparison, we consider BERT and BERT-BiLSTM-CRF as two baselines, which treat NER as a traditional sequence labeling problem. For the parameter settings of all models, the max sequence length is 512. The batch size is 8. The learning rate is 5×10^{-5} . The number of layers, neurons, and dropout ratio in BiLSTM is 1, 384, and 0.5, respectively. The epoch is 40, and the model with the best F1-score in the development set will be the adopted system.

4 Results and Discussion

Table 2 and Table 3 show the experimental results on flat NER and nested NER, respectively. As shown in Table 2, for flat NER, QASL-based models are slightly inferior to the baseline models by -0.39% (in terms of F1-score) for BERT-QA (vs. BERT) and by -1.16% for BERT-QA-BiLSTM-CRF (vs. BERT-BiLSTM-CRF). The slight decrease in performance of QASL-based models results from two main reasons: (1) QASL-based models are primarily designed to solve nested NER. Thus, QASL-based models are much more complicated than SL-based models, so that they are overqualified for flat NER that is far simpler than nested NER. (2) searching spaces of QASL-based models are much larger than that of SL-based models. QASL-based models are designed to search for various possible NEs without knowing how many they are in given passages in advance. In contrast, SL-based models directly assume that each possible entity span only has one entity type. The above two reasons cause the slight decrease of F1-score of QASL-based models compared to SL-based models.

As shown in Table 3, for nested NER, we observed that QASL-based models significantly

Model	P	R	F1
Bert	89.39	78.83	83.78
-BiLSTM-CRF	89.02	78.74	83.56
Bert-QA	87.67	92.26	89.90
-BiLSTM-CRF	91.01	90.40	90.70

Table 3: Model Performance on nested NER.

outperformed baseline models by +6.12% and +7.14% for BERT-QA (vs. BERT) BERTQA-BiLSTM-CRF (vs. BERT-BiLSTM-CRF), respectively. The substantial improvement of F1-scores is mainly from the boosted recall scores, attributed to the framework of QASL, which successfully detects nested NEs in the given queries and passages. Additionally, BERT-QA-BiLSTM-CRF achieves a 90.70% F1-score, which is +0.80% over that of BERT-QA. This is primarily because the BiLSTM-CRF structure makes QASL-based models assign more reasonable labels to tokens, reducing impossible outputs, thus leading to a higher F1-score.

5 Related Work

5.1 Named Entity Recognition

Most traditional feature-based approaches treated NER as a sequence labeling problem, thereby adopting Conditional Random Field (CRF) to resolve the NER task (Lafferty et al., 2001; Ratnikov and Roth, 2009). Recently, deep learning techniques have achieved good results on NER tasks, such as LSTM (Hammerston, 2003), CNN-CRF (Collobert et al., 2011), BiLSTM-CRF (Huang et al., 2015), and BiLSTM-CNN-CRF (Ma and Hovy, 2016). Besides, transfer learning has been applied to language models to improve model performance, such as ELMo (Peters et al., 2018), and BERT (Devlin et al., 2019). However, nested named entities cannot be recognized by the above approaches.

5.2 Nested Named Entity Recognition

Stack-based approaches have been used to extract entities from inner to outer or outer to inner, can handle the nested NER task. Alex et al. (Alex et al., 2007) proposed two multi-layers CRF models to recognize nested named entities; however, this approach cannot handle nested entities of the same entity type. Ju et

al. first (Ju et al., 2018) introduced a layered sequence labeling model to recognize innermost entities and then feed them into the next layer to extract outer entities. This method can deal with nested entities of the same type but suffers from error propagation among layers. Wang et al. (Wang et al., 2020a) proposed Pyramid, a novel layered model consisting of a stack of interconnected layers, to recognize entities without layer disorientation and error propagation.

Graph-based approaches have also been proposed to solve the nested NER task. Finkel and Manning (Finkel and Manning, 2009) used a CRF-based model to detect nested named entities with the assistance of constituency parse trees. Lu and Roth (Lu and Roth, 2015) introduced a hypergraph allowing edges to connect to multiple nodes to recognize overlapping entities. Wang and Lu (Wang and Lu, 2018) improved the spurious structures of the hypergraph by proposing neural segmental hypergraphs. Katiyar and Cardie (Katiyar and Cardie, 2018) used a LSTM model to learn a hypergraph representation for nested named entities. However, the hypergraph structure would become too complicated to be optimized if there are too many entities in the input sentences. Luo et al. (Luo and Zhao, 2020) proposed a novel bipartite flat graph network to recognize outermost entities and then use a graph module to extract inner ones.

Region-based approaches have utilized a pipeline framework with an end-to-end training paradigm to resolve the nested NER task. Specifically, these approaches first extract possible spans from the input sentence and then classify their entity types. Xu et al. (Xu et al., 2017) examined all possible spans (up to a certain length) of the input sentence and then fed their representation into a feed-forward neural network to classify entity types. Fisher and Vlachos 2019 (Fisher and Vlachos, 2019) first merged tokens into entities through real-valued predictions and then labeled them the corresponding entity types. Xia et al., 2019 (Xia et al., 2019) detected all possible spans through a detector and classified entities into pre-defined categories. Zheng et al., 2019 (Zheng et al., 2019) applied a single-layer sequence labeling model to identify the bound-

aries of potential entities using context information and then classify these boundary-aware regions into their entity type or non-entity. Wang et al., 2020 (Wang et al., 2020b) developed a head-tail detector and a token interaction tagger to identify nested named entities with appropriate model complexity.

Some researchers have attempted to transform NLP tasks into QA tasks, such as relation extraction (Levy et al., 2017; Li et al., 2019), summarization (McCann et al., 2018), named entity recognition (Li et al., 2020), and sentiment analysis (Yin et al., 2020). Li et al., 2020 (Li et al., 2020) treated NER as a QA problem. Each entity (y) and its entity type (x) can be parameterized as a question ($q(x)$) whose answer is (y). According to questions, models can parallelly identify nested named entities by using different questions. In addition, they can naturally solve flat NER as well.

6 Conclusion

This paper proposes a novel QA-based sequence labeling (QASL) approach to solve both flat and nested NER. The proposed framework comes with three key advantages: (1) It can recognize both flat and nested entities with a single model; (2) It combines QA and SL framework to solve NER and the problem of multiple spans selection; (3) The queries, encoding significant prior knowledge about entity types, are constructed without manual cost and are independent. The conducted experiments on Chinese electronic health records (CEHRs) have clearly shown the effectiveness of our proposed framework.

Acknowledgments

The research was supported by Data Intelligence R&D Division in CTBC Bank. We also thank Wen-Hsin Tseng for the assistance with data annotation.

References

Beatrice Alex, Barry Haddow, and Claire Grover. 2007. Recognising nested named entities in biomedical text. In *Biological, translational, and clinical language processing*, pages 65–72, Prague, Czech Republic. Association for Computational Linguistics.

- Julie Medero Christopher Walker, Stephanie Strassel, and Kazuaki Maeda. 2006. Ace 2005 multilingual training corpus. Philadelphia 57. Linguistic Data Consortium.
- Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of machine learning research*, 12(ARTICLE):2493–2537.
- Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Ziqing Yang, Shijin Wang, and Guoping Hu. 2019. Pre-training with whole word masking for chinese bert. *arXiv preprint arXiv:1906.08101*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. The automatic content extraction (ACE) program – tasks, data, and evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC’04)*, Lisbon, Portugal. European Language Resources Association (ELRA).
- Jenny Rose Finkel and Christopher D. Manning. 2009. Nested named entity recognition. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 141–150, Singapore. Association for Computational Linguistics.
- Joseph Fisher and Andreas Vlachos. 2019. Merge and label: A novel neural network architecture for nested NER. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5840–5850, Florence, Italy. Association for Computational Linguistics.
- James Hammerton. 2003. Named entity recognition with long short-term memory. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 172–175.
- Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.
- Meizhi Ju, Makoto Miwa, and Sophia Ananiadou. 2018. A neural layered model for nested named entity recognition. 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 1446–1459, New Orleans, Louisiana. Association for Computational Linguistics.
- Arzoo Katiyar and Claire Cardie. 2018. Nested named entity recognition revisited. 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 861–871, New Orleans, Louisiana. Association for Computational Linguistics.
- J-D Kim, Tomoko Ohta, Yuka Tateisi, and Jun’ichi Tsujii. 2003. Genia corpus—a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(suppl_1):i180–i182.
- John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.
- 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, San Diego, California. Association for Computational Linguistics.
- Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. 2017. Zero-shot relation extraction via reading comprehension. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 333–342, Vancouver, Canada. Association for Computational Linguistics.
- Xiaoya Li, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, and Jiwei Li. 2020. A unified MRC framework for named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5849–5859, Online. Association for Computational Linguistics.
- Xiaoya Li, Fan Yin, Zijun Sun, Xiayu Li, Arianna Yuan, Duo Chai, Mingxin Zhou, and Jiwei Li. 2019. Entity-relation extraction as multi-turn question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1340–1350, Florence, Italy. Association for Computational Linguistics.
- Wei Lu and Dan Roth. 2015. Joint mention extraction and classification with mention hypergraphs. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 857–867, Lisbon, Portugal. Association for Computational Linguistics.

- Ying Luo and Hai Zhao. 2020. Bipartite flat-graph network for nested named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6408–6418, Online. 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, Berlin, Germany. Association for Computational Linguistics.
- Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2018. The natural language decathlon: Multitask learning as question answering. *arXiv preprint arXiv:1806.08730*.
- Matthew E. 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, New Orleans, Louisiana. Association for Computational Linguistics.
- Lance A Ramshaw and Mitchell P Marcus. 1999. Text chunking using transformation-based learning. In *Natural language processing using very large corpora*, pages 157–176. Springer.
- 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-2009)*, pages 147–155, Boulder, Colorado. Association for Computational Linguistics.
- Nicky Ringland, Xiang Dai, Ben Hachey, Sarvnaz Karimi, Cecile Paris, and James R. Curran. 2019. NNE: A dataset for nested named entity recognition in English newswire. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5176–5181, Florence, Italy. Association for Computational Linguistics.
- Elad Segal, Avia Efrat, Mor Shoham, Amir Globerson, and Jonathan Berant. 2020. A simple and effective model for answering multi-span questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3074–3080, Online. Association for Computational Linguistics.
- Andrew Viterbi. 1967. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE transactions on Information Theory*, 13(2):260–269.
- Bailin Wang and Wei Lu. 2018. Neural segmental hypergraphs for overlapping mention recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 204–214, Brussels, Belgium. Association for Computational Linguistics.
- Jue Wang, Lidan Shou, Ke Chen, and Gang Chen. 2020a. Pyramid: A layered model for nested named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5918–5928, Online. Association for Computational Linguistics.
- Yu Wang, Yun Li, Hanghang Tong, and Ziyi Zhu. 2020b. HIT: Nested named entity recognition via head-tail pair and token interaction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6027–6036, Online. Association for Computational Linguistics.
- Congying Xia, Chenwei Zhang, Tao Yang, Yaliang Li, Nan Du, Xian Wu, Wei Fan, Fenglong Ma, and Philip Yu. 2019. Multi-grained named entity recognition. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1430–1440, Florence, Italy. Association for Computational Linguistics.
- 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)*, pages 1237–1247, Vancouver, Canada. Association for Computational Linguistics.
- Da Yin, Tao Meng, and Kai-Wei Chang. 2020. SentiBERT: A transferable transformer-based architecture for compositional sentiment semantics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3695–3706, Online. Association for Computational Linguistics.
- Changmeng Zheng, Yi Cai, Jingyun Xu, Ho-fung Leung, and Guandong Xu. 2019. A boundary-aware neural model for nested named entity recognition. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 357–366, Hong Kong, China. Association for Computational Linguistics.