Review of Neural Network-based Named Entity Recognition Methods

Shuyao Zhou
University of California, Berkeley
shuyaozhou@berkeley.edu

May 9, 2022

Abstract

Named Entity Recognition (NER) is a sub-task of information extraction. NER is widely applied in question answering, information retrieval, co-reference, topic modeling, machine translation, etc. (Yadav and Bethard, 2018). This paper[1] first introduces the history of NER, with Bidirectional Encoder Representations from Transformers (BERT) as the cutoff, and then presents the state-of-the-art neural network approaches.

1 Introduction

A name entity is a word or a phrase that identifies one item, such as organization, person, or location, according to predefined entity categories (Li et al., 2022). Named Entity Recognition (NER) was first proposed at the Sixth Message Understanding Conference in 1996 (Grishman and Sundheim, 1996), which is a sub-task of information extraction, identifying, and classifying named entities in texts. More formally, given an input sequence x, a label y from the set of valid NER labels, is assigned to each of the spans in $S(x)$.

1.1 Conventional NER

Before the extensive use of neural networks in NER, classical approaches, such as the Hidden Markov Model (HMM), Maximum Entropy Markov Model (MEMM), and Conditional Random Fields (CRF), were commonly used. These methods are still widely accepted now. Also, researchers take sequence structure, or the previous classified words, into account to make the tag decision for the following word. Ratinov and Roth (2009) make comprehensive comparisons among conventional NER approaches. They identify some fundamental design challenges and misconceptions of NER methods. For example, they illustrate that BILOU representation of text chunks performs better than BIO. Their work acts as a guideline for future research.

1.2 Mainstream NER Datasets

Creating NER datasets is a major challenge. CoNLL2003 (Tjong Kim Sang and De Meulder, 2003) is one of the benchmark NER datasets. It is a language-independent NER dataset, with English and German as training and test data, that focuses on four types of named entities: persons (PER), locations (LOC), organizations (ORG), and names of miscellaneous entities (MISC). The English data was taken from Reuters Corpus, and the German data was taken from the ECI Multilingual Text Corpus. Based on CoNLL2003, CrossWeigh (Wang et al., 2019) corrected some label mistakes in the CoNLL2003 test set, and a corresponding framework CorssWeigh is proposed to handle label mistakes during training. Weischedel et al. (2013) created OntoNotes 5.0 to annotate a large corpus of newswire (News), broadcast news (BN), broadcast conversation (BC), telephone conversation (Tele), and web data (Web) in English, Chinese, and Arabic. Wnut 16 (Strauss et al., 2016) is a NER dataset on Twitter, which is challenging because the information on social media is noisy. Additionally, there are many multi-lingual datasets being developed such as (Bari et al., 2019) and (Al-Rfou et al., 2015).

1.3 Neural Network-based NER

Neural networks have been the state-of-the-art NER approaches. Therefore, this survey will focus on neural network-based methods tackling entity linking, low-resource learning, and general robustness. Hammerton (2003) was the first to use Long Short Term Memory (LSTM) for NER. In 2011, Collobert et al. (2011) designed the CNN-CRF structure, obtaining competitive results. In

[1] Word Count: 2013

2015, Huang et al. (2015) proposed a Bidirectional LSTM-CRF model, which can efficiently utilize past and future input features. With the CRF layer, it can also use sentence-level tokenization information. Also, the model is robust and less dependent on word embeddings than previous models.

Lample et al. (2016) developed another classic model using stack-LSTM and char-embedding. They add a stack pointer to LSTM, and there are three actions: SHIFT (moves a word from the buffer to the stack), OUT (moves a word from the buffer to the output), and REDUCE (pops all the words in the stack into a block, tags them with tag y, and pushes them into output). During the training, the model gets the conditional probability distribution of each action step, and the label is the real probability of each action. For prediction, by predicting the probability of each action step, the model executed the highest probability action. After the REDUCE action, the vector representation of the chunk is encoded by LSTM and then its label is predicted.

To make NER datasets more accessible, Dernoncourt et al. (2017) build NeuroNER based on artificial neural networks (ANN). Users can use a graphical web-based interface to annotate entities, and these annotations can then be used to train an ANN, which in turn can be used to predict the location and class of entities in new texts. Their work enables not only the experts but also anyone, to create or modify annotations for a new or existing corpus.

One drawback of traditional NER is that the segment error propagates backward when the entity segmentation is wrong, making NER generate fatal errors. To solve this, Zhang and Yang proposed a Lattice LSTM (Zhang and Yang, 2018) to automatically control the information flow from the beginning to the end of the input sequence using LSTM because the lattice partitioning grows exponentially (See Figure 1). With the introduction of BERT (Devlin et al., 2019), NER has come to a new stage. Besides sequence labeling, there are more and more methods being neural-based.

2 Neural Network-based NER

2.1 Entity Linking

Entity linking (EL), or entity alignment, is the task of linking entities with the equivalent meaning from different knowledge graphs (KGs). KGs store entities and their relationships using graph structures (See Figure 2). Neural-based methods are state-of-the-art in EL, but they often require seed alignments as training data.

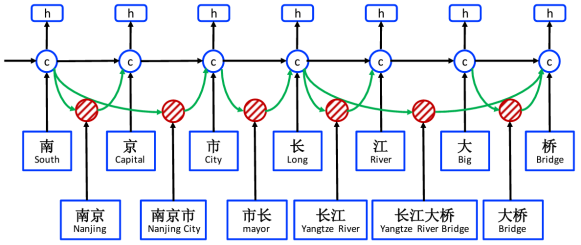

Figure 1: Lattice LSTM structure (Zhang and Yang, 2018).

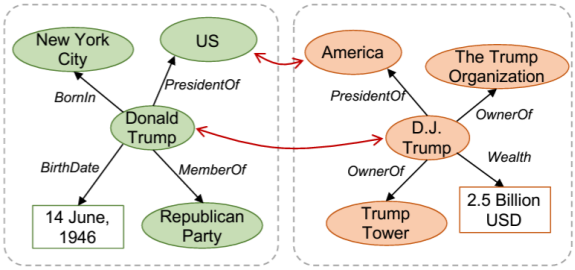

Figure 2: An example of Entity Linking. "Donald Trump" and "US" in the left KG correspond to "D.J. Trump" and "America" respectively in the right KG (Liu et al., 2021).

In real life, not every domain has labeled data available but has its own KGs. Recently, researchers started to investigate unsupervised EL that can operate with access only to KGs and entity names but not to annotated data. In 2019, Logeswaran et al. (2019) presented the zero-shot entity linking task and made a dataset for it. In 2021, Arora et al. (2021) proposed an unsupervised framework called EIGENTHEMES. They assumed that the entities mentioned in a document, or gold entities, tend to lie in a low-rank subspace of the full embedding space. EIGENTHEMES uses singular value decomposition to learn the subspace spanned by a certain hyper-parameter number of components where each principal component captures the topical relatedness among gold entities across different contexts.

Liu et al. (2021) proposed active learning for neural entity alignment (ActiveEA) to reduce the cost of annotation for seed alignment training data. Active Learning (AL) means that a learning algorithm can iteratively choose the data that is going to be annotated (Settles, 2009). After the query system selects the entities to be annotated, they

define the influence of an entity on its context as the number of uncertainties it can help its neighbors remove. In each iteration, structure-aware uncertainty sampling will combine the structure information of KGs with uncertainty sampling, and the bachelor recognizer will reduce the annotation cost by avoiding the selection of bachelors. They were the first to exploit active learning in EL to obtain the most informative entities for labeling.

On top of the neural-based models, Yang et al. (2019) enhanced the learning process by applying the dynamic context augmentation process in ETHZ-Attn model (Ganea and Hofmann, 2017) and Berkeley-CNN (Francis-Landau et al., 2016) model. A traditional global EL model iteratively calculates all mentions, and then jointly optimizes the linking. Requiring less computation, the DCA only requires one pass of the document to accumulate knowledge from previously linked mentions for future reference.

2.2 Low-resource Learning

Neural NER models require large labeled training datasets, but in most cases, the labeled datasets are small, and entity categories are not always the same in different domains. Another problem is that both softmax and CRF require consistent labels between training and testing. Therefore, there is a growing emphasis on the few-shot NER, which is performing labeling tasks when labeled examples are scarce.

Bidirectional and auto-regressive transformers (BART) (Lewis et al., 2020) is a denoising autoencoder for pretraining sequence-to-sequence models. Using BART as the backbone, Cui et al. (2021) was the first to apply template-based methods in few-shot NER. Within the sequence-to-sequence framework, NER is considered a language model ranking problem, where the input texts are the source sequences, and the templates filled by candidate text span and the named entity span are the target sequences. For example, given the sentence "James is in Sydney", where "Sydney" has a gold label "location". Then, they train BART using a filled template "Sydney is a location entity" as the decoder output for the input sentence.

Huang et al. (2021) studied three orthogonal strategies to improve the generalization ability of few-shot NER: prototype methods, noisy supervised pre-training, and self-training. They claim that noisy supervised pre-training can significantly improve NER accuracy. They also present the first study of self-training for NER, which can improve few-shot learning. Moreover, prototype learning, which utilizes the nearest-neighbor criterion to assign the entity type by comparing their prototypes, performs better when the number of labeled examples is small.

Unsupervised consistency training encourages consistency in model predictions between the original data and augmented data. In 2020, Xie et al. (2020) proposed unsupervised data augmentation (UDA), which is a semi-supervised learning approach that does not require large labeled datasets and makes better use of unlabeled data. UDA substitutes noising operations with data augmentation methods such as paraphrasing via back-translation for consistency training. For example, the original text is "Given the low budget and production limitations, this movie is very good". After back-translation, the text becomes "Due to small dollar amount and production limitations the ouest film is very beautiful". Later, Lowell et al. (2021) randomly replaced input tokens with model predictions and enforced consistency between predictions assigned to observe these randomly substituted words. They illustrated that UDA did not require complex data augmentation, such as back-translation, to be effective. Based on these, Wang and Henao (2021) explored the use of paraphrasing for higher quality data augmentation for unsupervised consistency training in NER. Instead of exploring the token-level consistency, they focus on consistency in the occurrence of entities between the original sequence and the prediction of the paraphrased sequence. Their model consists of a BERT-based encoder and a CRF module for prediction.

2.3 Robustness

Increasing efficiency and accuracy while maintaining stability for new datasets is a mainstream topic for NER tasks.

Entity tagging is a lasting topic in NER tasks. For example, BIO is a classic approach. A recurrent neural network (RNN) is a neural network that feeds the output from the previous step to the current step while having hidden states, and it can handle arbitrary input lengths. Bidirectional long short-term memory (BiLSTM) has two networks with one accessing past information in the forward direction and another accessing future in the re-

verse direction. CRF predicts the sequence using contextual information. Li et al. (2021c) used an RNN encoder-decoder framework with a pointer network to detect entity segments. In 2021, Li et al. (2021b) proposed a modularized interaction network model. They designed three modules consisting of the NER Module, Boundary Module, and Type Module (See Figure 3). In the NER Module, they adopted the RNN-BiLSTM-CRF model: word representation, BiLSTM encoder, and CRF decode, to extract character-level word segment information. They also use a gate function to control the amount of information flowing by infusing the expedient part. In the Boundary Module, they use another BiLSTM encoder to extract distinct contextual boundary information. In the Type Module, they use the RNN-BiLSTM-CRF model again. Given the shared input, BiLSTM extracts distinct contextual type information, and CRF tags type labels.

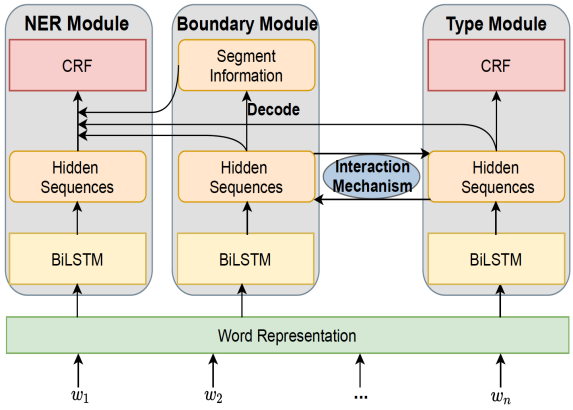

Figure 3: Modularized Interaction Network architecture (Zhang and Yang, 2018).

Additionally, boundary tagging ambiguity raises attention from researchers. Boundary tag sparsity occurs when entities are rare in a sentence. Recently, Li et al. (2021a) tackle the boundary tag sparsity problem with a boundary-aware bidirectional neural network model, which is based on an encoder-decoder frame and integrated with the pointer network. They also try to capture the global decoding information and alleviate boundary error propagation problems with the model and achieve state-of-the-art performance.

For evaluating the robustness of NER, RockNER (Lin et al., 2021), proposed by Lin et al., can create natural adversarial examples. Their pipeline consists of two steps: at the entity-level attack, they substitute target entities with other entities of the same semantic class in Wikidata; at the context-level attack, they exploit pre-trained language models such as BERT to substitute other target entities. The two levels of attack produce adversarial examples that can shift the distribution of the training data on the evaluated model. They recreate the OntoNotes dataset to OntoRock, and the new benchmark shows that the state-of-the-art models have a significant performance drop. These evaluated models tend to not reason from the contexts. RockNER can be a metric to improve the robustness and data augmentation of NER models for future research. Similarly, SeqAttack (Simoncini and Spanakis, 2021) also proposed adversarial attacks working against NER models and for data augmentation.

3 Conclusion

NER is useful in natural language tasks involving entities, and researchers have worked on NER with high efficiency and stability. The purpose of this paper is to highlight the milestones for NER approaches before BERT, and the current state-of-the-art subfields after BERT, which are entity linking, low resource learning, and robustness. There are other potential topics such as cross-domain learning, generalization, dataset creation, discontinuous NER, etc. that the research field grows continuously.

References

Rami Al-Rfou, Vivek Kulkarni, Bryan Perozzi, and Steven Skiena. 2015. Polyglot-ner: Massive multilingual named entity recognition. In *Proceedings of the 2015 SIAM International Conference on Data Mining*, pages 586–594. SIAM.

Akhil Arora, Alberto Garcia-Duran, and Robert West. 2021. Low-rank subspaces for unsupervised entity linking. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8037–8054, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

M. Saiful Bari, Shafiq R. Joty, and Prathyusha Jwalapuram. 2019. Zero-resource cross-lingual named entity recognition. *CoRR*, abs/1911.09812.

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(76):2493–2537.

Leyang Cui, Yu Wu, Jian Liu, Sen Yang, and Yue Zhang. 2021. Template-based named entity recognition using BART. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1835–1845, Online. Association for Computational Linguistics.

Franck Dernoncourt, Ji Young Lee, and Peter Szolovits. 2017. NeuroNER: an easy-to-use program for named-entity recognition based on neural networks. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 97–102, Copenhagen, Denmark. Association for Computational Linguistics.

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.

Matthew Francis-Landau, Greg Durrett, and Dan Klein. 2016. Capturing semantic similarity for entity linking with convolutional neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1256–1261, San Diego, California. Association for Computational Linguistics.

Octavian-Eugen Ganea and Thomas Hofmann. 2017. Deep joint entity disambiguation with local neural attention. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2619–2629, Copenhagen, Denmark. Association for Computational Linguistics.

Ralph Grishman and Beth Sundheim. 1996. Message Understanding Conference- 6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on 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.

Jiaxin Huang, Chunyuan Li, Krishan Subudhi, Damien Jose, Shobana Balakrishnan, Weizhu Chen, Baolin Peng, Jianfeng Gao, and Jiawei Han. 2021. Few-shot named entity recognition: An empirical baseline study. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10408–10423, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

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.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Fei Li, Zheng Wang, Siu Cheung Hui, Lejian Liao, Dandan Song, and Jing Xu. 2021a. Effective named entity recognition with boundary-aware bidirectional neural networks. In *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*, pages 1695–1703. ACM / IW3C2.

Fei Li, Zheng Wang, Siu Cheung Hui, Lejian Liao, Dandan Song, Jing Xu, Guoxiu He, and Meihuizi Jia. 2021b. Modularized interaction network for named entity recognition. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 200–209, Online. Association for Computational Linguistics.

Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2022. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, 34(1):50–70.

Jing Li, Aixin Sun, and Yukun Ma. 2021c. Neural named entity boundary detection. *IEEE Transactions on Knowledge and Data Engineering*, 33(4):1790–1795.

Bill Yuchen Lin, Wenyang Gao, Jun Yan, Ryan Moreno, and Xiang Ren. 2021. Rockner: A simple method to create adversarial examples for evaluating the robustness of named entity recognition models. In *Proc. of EMNLP (short paper)*. To appear.

Bing Liu, Harrisen Scells, Guido Zuccon, Wen Hua, and Genghong Zhao. 2021. ActiveEA: Active learning for neural entity alignment. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3364–3374, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Lajanugen Logeswaran, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, Jacob Devlin, and Honglak Lee. 2019. Zero-shot entity linking by reading entity descriptions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*,

pages 3449–3460, Florence, Italy. Association for Computational Linguistics.

David Lowell, Brian Howard, Zachary C. Lipton, and Byron Wallace. 2021. Unsupervised data augmentation with naive augmentation and without unlabeled data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4992–5001, Online and Punta Cana, Dominican Republic. 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-2009)*, pages 147–155, Boulder, Colorado. Association for Computational Linguistics.

Burr Settles. 2009. Active learning literature survey.

Walter Simoncini and Gerasimos Spanakis. 2021. SeqAttack: On adversarial attacks for named entity recognition. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 308–318, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Benjamin Strauss, Bethany Toma, Alan Ritter, Marie-Catherine de Marneffe, and Wei Xu. 2016. Results of the WNUT16 named entity recognition shared task. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pages 138–144, Osaka, Japan. The COLING 2016 Organizing Committee.

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*, pages 142–147.

Rui Wang and Ricardo Henao. 2021. Unsupervised paraphrasing consistency training for low resource named entity recognition. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5303–5308, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zihan Wang, Jingbo Shang, Liyuan Liu, Lihao Lu, Jiacheng Liu, and Jiawei Han. 2019. Crossweigh: Training named entity tagger from imperfect annotations. *arXiv preprint arXiv:1909.01441*.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, et al. 2013. Ontonotes release 5.0 ldc2013t19. *Linguistic Data Consortium, Philadelphia, PA*, 23.

Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. 2020. Unsupervised data augmentation for consistency training. In *Advances in Neural Information Processing Systems*, volume 33, pages 6256–6268. Curran Associates, Inc.

Vikas Yadav and Steven Bethard. 2018. A survey on recent advances in named entity recognition from deep learning models. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2145–2158, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Xiyuan Yang, Xiaotao Gu, Sheng Lin, Siliang Tang, Yueting Zhuang, Fei Wu, Zhigang Chen, Guoping Hu, and Xiang Ren. 2019. Learning dynamic context augmentation for global entity linking. 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 271–281, Hong Kong, China. Association for Computational Linguistics.

Yue Zhang and Jie Yang. 2018. Chinese NER using lattice LSTM. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1554–1564, Melbourne, Australia. Association for Computational Linguistics.