Gazetteer-Enhanced Attentive Neural Networks for Named Entity Recognition

Hongyu Lin[1,3], **Yaojie Lu**[1,3], **Xianpei Han**[1,2,*], **Le Sun**[1,2], **Bin Dong**[4], **Shanshan Jiang**[4]
[1]Chinese Information Processing Laboratory [2]State Key Laboratory of Computer Science
Institute of Software, Chinese Academy of Sciences, Beijing, China
[3]University of Chinese Academy of Sciences, Beijing, China
[4]Ricoh Software Research Center (Beijing) Co.,Ltd.
{hongyu2016,yaojie2017,xianpei,sunle}@iscas.ac.cn
{bin.dong,shanshan.jiang}@srcb.ricoh.com

Abstract

Current region-based NER models only rely on fully-annotated training data to learn effective region encoder, which often face the training data bottleneck. To alleviate this problem, this paper proposes *Gazetteer-Enhanced Attentive Neural Networks*, which can enhance region-based NER by learning name knowledge of entity mentions from easily-obtainable gazetteers, rather than only from fully-annotated data. Specially, we first propose an *attentive neural network (ANN)*, which explicitly models the mention-context association and therefore is convenient for integrating externally-learned knowledge. Then we design an auxiliary *gazetteer network*, which can effectively encode name regularity of mentions only using gazetteers. Finally, the learned gazetteer network is incorporated into ANN for better NER. Experiments show that our ANN can achieve the state-of-the-art performance on ACE2005 named entity recognition benchmark. Besides, incorporating gazetteer network can further improve the performance and significantly reduce the requirement of training data.

1 Introduction

Named entity recognition (NER), aiming to identify text mentions of specific entity types, is a fundamental NLP task. Recently, region-based NER approaches (Finkel and Manning, 2009; Xu et al., 2017; Sohrab and Miwa, 2018) have attracted significant attention, which first encode all candidate regions (commonly all subsequences of a sentence) using a *region encoder*, then identify whether each subsequence is an entity mention of target types using a classifier. For example, in Figure 1 all subsequences of the sentence, such as "George Washington", will first be encoded, and then be classified into entity types. Compared to sequential labeling models, region-based models can naturally detect nested or overlapping mentions by considering all subsequences, and therefore are of great value to NER.

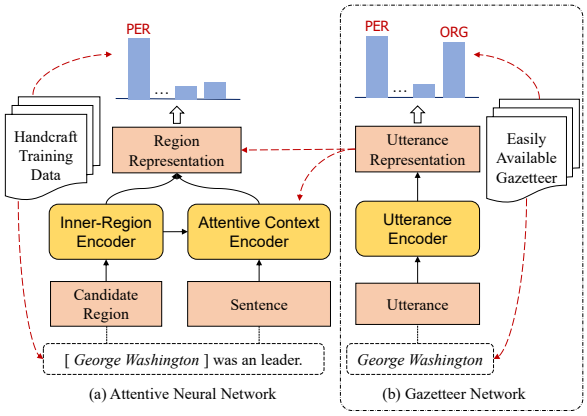

Figure 1: The overall architecture of GEANN. The candidate region is "George Washington", which literally could be a person or an organization (university).

Generally, an effective region encoder should capture two kinds of knowledge for NER. One is *name knowledge*, which encodes the inner compositional regularity of entity mentions, i.e., how likely a subsequence itself will be an entity mention. For example, a region encoder should know "George Washington" is a valid *PER* name because "George" is a common first name and "Washington" is a common last name. Another is *context knowledge*, which identifies whether the subsequence in the context indeed refers to an entity. For example, a region encoder should know "X said" is a suitable context for *PER* mention and "study at X" is a suitable context for *Org* mention.

Currently, most region-based NER models learn these two kinds of knowledge only from expensive, fully-annotated training data, and therefore often face the training data bottleneck, i.e., the lack of training data will significantly undermine

*Xianpei Han is the corresponding author.

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 6232–6237,
Hong Kong, China, November 3–7, 2019.

the performance. To address this problem, we find that name knowledge can be effectively captured by leveraging easily-obtainable gazetteer resources. For example, it is easy to learn the company name patterns "the...company" and "...Inc." from a company name gazetteer containing "the Walt Disney company" and "Apple Inc.". By capturing mention regularity entailing gazetteers, the region-based models can be enhanced with more accurate name knowledge, and thereby the need of fully-annotated training data can be reduced.

To this end, this paper proposes ***G**azetteer-**E**nhanced **A**ttentive **N**eural **N**etworks (GEANN)*, whose architecture is shown in Figure 1. Specifically, to better decouple name and context knowledge for incorporating gazetteer information, we first design a new region-based ***attentive neural network (ANN)***, which introduces attention mechanism (Bahdanau et al., 2014; Vaswani et al., 2017; Su et al., 2018) to explicitly model the association between mentions and contexts. Starting from ANN, we further introduce an auxiliary ***gazetteer network***, which can effectively learn name knowledge only using gazetteers, i.e., it can encode each utterance in a context-free way, and identifies whether it matches regular patterns of mentions. Finally, the learned gazetteer network is incorporated into ANN to capture better name and context knowledge. Experiments show that ANN can achieve the state-of-the-art NER performance, and incorporating name knowledge from gazetteers can significantly reduce the training data requirement. To the best of our knowledge, this is the first work trying to explicitly exploit mention-context association with attention mechanism in region-based NER, as well as the first work which enhances NER model with name knowledge captured from gazetteers using neural networks.

2 Attentive Neural Network for NER

This section describes our attentive neural network, which directly classifies over all subsequences of a sentence to recognize whether each subsequence corresponds to an entity mention. Figure 1 (a) shows the architecture of ANN.

Given a sentence, ANN first maps all words into word representations $\{\boldsymbol{x_1}, \boldsymbol{x_2}, ..., \boldsymbol{x_n}\}$ following Lample et al. (2016). Then a BiLSTM layer is used to obtain context-aware word representation $\boldsymbol{h_t^A}$. After that, for each candidate region s_{ij}, we follow Sohrab and Miwa (2018) to use an inner-region encoder to obtain its representation $\boldsymbol{s_{ij}}$, which captures name knowledge considering both its boundary and inside information:

$$\boldsymbol{s}_{ij} = \text{MLP}([\boldsymbol{h}_i^A; \frac{1}{j-i+1}\sum_{k=i}^{j}\boldsymbol{h}_k^A; \boldsymbol{h}_j^A]), \quad (1)$$

where MLP is a multi-layer perceptron. To explicitly model the association between a region and its context, we design an attentive context encoder, which outputs a contextual vector $\boldsymbol{c}_{ij}$ entailing the context knowledge of s_{ij} by:

$$\begin{aligned} \boldsymbol{c}_{ij} &= \sum_{k=1}^{i-1}\alpha_{ijk}\boldsymbol{h}_k^A + \sum_{k=j+1}^{n}\alpha_{ijk}\boldsymbol{h}_k^A \\ \alpha_{ijk} &= \frac{\exp(e_{ijk})}{\sum_{t=1}^{i-1}\exp(e_{ijt}) + \sum_{t=j+1}^{n}\exp(e_{ijt})} \end{aligned} \quad (2)$$

where

$$e_{ijk} = \tanh(\boldsymbol{s_{ij}}^T\boldsymbol{\Lambda}\boldsymbol{h}_k^A + \boldsymbol{W}\boldsymbol{h}_k^A + b) \quad (3)$$

is an attentive model which scores how important the word x_k is for recognizing the entity type of s_{ij}. After obtaining $\boldsymbol{c}_{ij}$, we concatenate it with $\boldsymbol{s}_{ij}$, and feed it into a MLP classifier to obtain the probabilities of s_{ij} corresponding to an entity type (or NIL if s_{ij} is not a mention). Similar to previous methods, ANN is learned by minimizing the cross-entropy loss on the fully-annotated training data.

By decoupling and explicitly modeling the mention-context association, ANN not only can better identify entity mentions, but also is very easy to incorporate external name knowledge. This enables convenient integration of gazetteer knowledge, as we will illustrate next.

3 Gazetteer-Enhanced ANN

One main drawback of ANN and other region-based models is that they only rely on fully-annotated data to learn both name knowledge and context knowledge. Unfortunately, such training data is very expensive to construct, which limits the applications of these model to more entity types. To tackle this problem, we proposes to learn name knowledge from easily-obtainable, large-scale gazetteers. In this way, the name knowledge can be captured more accurately, and the requirement of fully-annotated data can be reduced.

To this end, we propose to incorporate an auxiliary *gazetteer network* into ANN, which can learn and exploit name knowledge only using gazetteers. Given an utterance, gazetteer network predicts

whether it should be included in the gazetteer of specific entity types, i.e., whether it is a valid entity name. Gazetteer network is context-free, which only considers whether the input follows the compositional regularity of mentions, and therefore can be trained only using gazetteers.

Formally, given an input utterance $u = \{u_i, ..., u_j\}$, an utterance encoder first learns its representation $\boldsymbol{u}$, which has similar structure with the inner-region encoder of ANN. After that, $\boldsymbol{u}$ is used to compute the probabilities of it being a valid name for each type:

$$\boldsymbol{O}_u^G = s(\boldsymbol{W}\boldsymbol{u} + \boldsymbol{b}). \quad (4)$$

where s is the sigmoid function and k^{th} dimension of $\boldsymbol{O}_u^G$ indicates the probability of u being a valid mention of type y_k. As an utterance can possibly be a valid mention of various types, we use a multi-label, multi-class cross-entropy loss to train our gazetteer network:

$$\mathcal{L}^G(\theta) = \sum_{u \in G} [\boldsymbol{g}'_u \log \boldsymbol{O}_u^G + (1 - \boldsymbol{g}'_u) \log(1 - \boldsymbol{O}_u^G)] \quad (5)$$

where G is the gazetteers, $\boldsymbol{g}'_u$ is an one-hot vector whose k^{th} will be set to 1 if utterance u is in the gazetteer of type y_k, elsewise 0.

In this way, a well-trained gazetteer network can learn an effective utterance representation $\boldsymbol{u}$, which is used to identify whether the utterance is a valid mention. This means $\boldsymbol{u}$ should capture enough name knowledge of specific entity types. To incorporate such knowledge into ANN, we simply concatenate the representation learned by gazetteer network and the representation learned by the original inner-region encoder. Then this new representation is fed to the following modules of ANN. In this way, name knowledge learned from gazetteers is incorporated to enhance the region encoder, and the requirement of fully-annotated data can be reduced.

4 Experiments

4.1 Experimental Settings

Data Preparation. We conducted experiments on ACE2005 named entity recognition task[1]. We used the same dataset splits as Wang and Lu (2018); Katiyar and Cardie (2018). For each entity type in ACE2005, we collect a gazetteer from Wikipedia anchor texts, i.e., anchor texts linking to an entity whose type are the same will be included in a gazetteer. The same as previous studies, models are evaluated using micro-F1. To balance time complexity and recall, we follow Wang and Lu (2018) to restrict mention length up to 6, which covers more than 93% mentions.

[1]Conventional NER datasets, such as CoNLL2003, removed nested entity mentions and therefore are not suitable for evaluating region-based NER models.

	P	R	F1
LSTM-CRF	73.2	61.3	66.7
Neural Transition	75.2	65.5	70.0
Segmental Hypergraph	75.7	68.3	71.8
Exhaustive	**81.2**	66.9	73.4
ANN	78.9	69.8	74.1
GEANN	77.1	**73.3**	**75.2**

Table 1: Experiment results on ACE2005 named entity mention recognition.

Baselines. Following methods were compared:

1) **LSTM-CRF** (Lample et al., 2016), which is the most widely used NER baseline, but it cannot handle nested or overlapping mentions.

2) **Neural Transition** (Wang et al., 2018), a transition model which achieved very competitive performance on ACE2005.

3) **Segmental Hypergraph** (Wang and Lu, 2018), a hypergraph-based model which introduces a new tagging schema and achieved the state-of-the-art performance on ACE2005.

4) **Exhaustive Model** (Sohrab and Miwa, 2018), a region-based model using a region encoder to capture both inner and boundary features of a candidate region, which is similar to ANN without attentive contextual encoder.

4.2 Overall Results

Table 1 shows the overall results of our methods compared with baselines. We can see that:

1) By explicitly modeling both name and context knowledge, the proposed attentive neural network is an effective region-based NER model, and achieves the state-of-the-art performance. Compared with other baselines, ANN achieves significant F1-score improvements.

2) Incorporating name knowledge from gazetteers can significantly improve the performance. GEANN further achieves 1.1 F1-score improvement over ANN. This indicates that name knowledge learned from gazetteers can significant enhance the region encoder, and therefore improves the NER performance.

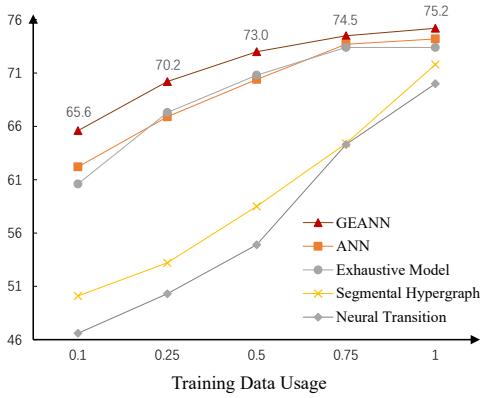

Figure 2: F1-scores on ACE2005 when training data size varies.

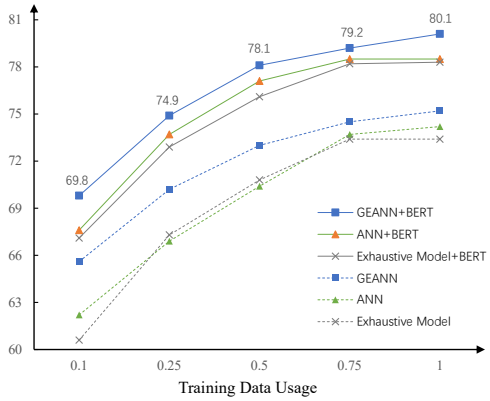

Figure 3: F1-scores of models with BERT on ACE2005 when training data size varies.

3) Our attentive context encoder provides an effective way to exploit context knowledge for NER. Compared with *Exhaustive* baseline, ANN achieves significant improvement by explicitly modeling the association between entity mentions and their contexts.

4.3 Effects of Gazetteer Network

To further investigate the effect of introducing gazetteers, Table 2 shows the results when training data size varies. We can see that:

1) For *Transition* and *SH*, their performance significantly decreased when reducing training data. We believe this because these approaches need to model complex label structure, so large scale training data are critical, and reducing training data will have huge impact on them.

2) Region-based models are less sensitive to the reduction of training data. We believe this is because their output structure is simple, and therefore can be trained using less training data.

3) GEANN can achieve significant improvements over ANN regardless of training data size. By leveraging name knowledge from gazetteers, GEANN with only 50% training data can achieve comparable performance with ANN trained with the entire dataset.

4.4 GEANN with BERT

Pretrained context-aware representation, such as ELMOs (Peters et al., 2018) and BERT (Devlin et al., 2018), have shown significant progress in many NLP tasks, especially in low-resource cases. To verify the adaptivity of the proposed GEANN, we further introduce BERT into models by replacing the word embedding with BERT representations. Figure 3 shows the results. We can see that even BERT can enhance NER, GEANN can still further achieve significant improvement over BERT regardless of the training data size. This verified GEANN can further capture task-specific name knowledge, which complement well with universal pretrained language knowledge.

5 Related Work

Sequential labeling approaches (Zhou and Su, 2002; Chieu and Ng, 2002; Bender et al., 2003; Settles, 2004; Lample et al., 2016) are widely used in NER. But this paradigm cannot handle nested mentions without specially designed tagging schema (Lu and Roth, 2015; Katiyar and Cardie, 2018; Wang and Lu, 2018; Lin et al., 2019). Recently, region-based models provide a natural solution for this issue. Finkel and Manning (2009) first proposed to classify over regions corresponding to parsing tree nodes. Xu et al. (2017) proposed to directly classify over all subsequences with a neural network model. Sohrab and Miwa (2018) extended their method by introducing a new region encoder. Generally, these methods have achieved promising results but heavily rely on fully-annotated data.

Gazetteers or dictionaries have long been regarded as a useful and easily-obtainable resource for NER. Previous methods commonly incorporated gazetteers by either using them to as handcraft features (Bender et al., 2003; Tsuruoka and Tsujii, 2003; Ciaramita and Altun, 2005; Minkov et al., 2005; Ritter et al., 2011; Seyler et al., 2018; Yu et al., 2018), or using them to generate data by distant supervision (Cohen, 2005; Ren et al., 2015; Giannakopoulos et al., 2017; Shang et al., 2018). However, the first kind of methods can not fully

leverage the inner mention structure knowledge entailing in gazetteers, while the second approaches will result in remarkable noise.

6 Conclusions

This paper first proposes *attentive neural networks*, an effective region-based model which explicitly models mention-context association. Then we propose to incorporate an auxiliary *gazetteer network* to enhance ANN. The gazetteer network can effectively learn name knowledge only using easily-available gazetteers, and therefore can significantly improve model performance and reduce data requirement. Experiments show that GEANN achieves the state-of-the-art performance on ACE2005 with much lower data requirement.

Acknowledgments

We sincerely thank the reviewers for their insightful comments and valuable suggestions. Moreover, this work is supported by the National Natural Science Foundation of China under Grants no. 61433015, 61572477 and 61772505; and the Young Elite Scientists Sponsorship Program no. YESS20160177.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Oliver Bender, Franz Josef Och, and Hermann Ney. 2003. Maximum entropy models for named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 148–151. Association for Computational Linguistics.

Hai Leong Chieu and Hwee Tou Ng. 2002. Named entity recognition: a maximum entropy approach using global information. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, pages 1–7. Association for Computational Linguistics.

Massimiliano Ciaramita and Yasemin Altun. 2005. Named-entity recognition in novel domains with external lexical knowledge. In *Proceedings of the NIPS Workshop on Advances in Structured Learning for Text and Speech Processing*, volume 2005.

Aaron M Cohen. 2005. Unsupervised gene/protein named entity normalization using automatically extracted dictionaries. In *Proceedings of the acl-ismb workshop on linking biological literature, ontologies and databases: Mining biological semantics*, pages 17–24. 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. *arXiv preprint arXiv:1810.04805*.

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: Volume 1-Volume 1*, pages 141–150. Association for Computational Linguistics.

Athanasios Giannakopoulos, Claudiu Musat, Andreea Hossmann, and Michael Baeriswyl. 2017. Unsupervised aspect term extraction with b-lstm & crf using automatically labelled datasets. *arXiv preprint arXiv:1709.05094*.

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. Association for Computational Linguistics.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*.

Hongyu Lin, Yaojie Lu, Xianpei Han, and Le Sun. 2019. Sequence-to-nuggets: Nested entity mention detection via anchor-region networks. *arXiv preprint arXiv:1906.03783*.

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.

Einat Minkov, Richard C Wang, and William W Cohen. 2005. Extracting personal names from email: Applying named entity recognition to informal text. In *Proceedings of human language technology conference and conference on empirical methods in natural language processing*.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Xiang Ren, Ahmed El-Kishky, Chi Wang, Fangbo Tao, Clare R Voss, and Jiawei Han. 2015. Clustype: Effective entity recognition and typing by relation phrase-based clustering. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 995–1004. ACM.

Alan Ritter, Sam Clark, Oren Etzioni, et al. 2011. Named entity recognition in tweets: an experimental study. In *Proceedings of the conference on empirical methods in natural language processing*, pages 1524–1534. Association for Computational Linguistics.

Burr Settles. 2004. Biomedical named entity recognition using conditional random fields and rich feature sets. In *Proceedings of the international joint workshop on natural language processing in biomedicine and its applications*, pages 104–107. Association for Computational Linguistics.

Dominic Seyler, Tatiana Dembelova, Luciano Del Corro, Johannes Hoffart, and Gerhard Weikum. 2018. A study of the importance of external knowledge in the named entity recognition task. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 241–246.

Jingbo Shang, Liyuan Liu, Xiang Ren, Xiaotao Gu, Teng Ren, and Jiawei Han. 2018. Learning named entity tagger using domain-specific dictionary. *arXiv preprint arXiv:1809.03599*.

Mohammad Golam Sohrab and Makoto Miwa. 2018. Deep exhaustive model for nested named entity recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2843–2849. Association for Computational Linguistics.

Jinsong Su, Jiali Zeng, Deyi Xiong, Yang Liu, Mingxuan Wang, and Jun Xie. 2018. A hierarchy-to-sequence attentional neural machine translation model. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 26(3):623–632.

Yoshimasa Tsuruoka and Jun'ichi Tsujii. 2003. Boosting precision and recall of dictionary-based protein name recognition. In *Proceedings of the ACL 2003 workshop on Natural language processing in biomedicine-Volume 13*, pages 41–48. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

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. Association for Computational Linguistics.

Bailin Wang, Wei Lu, Yu Wang, and Hongxia Jin. 2018. A neural transition-based model for nested mention recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1011–1017. 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. Association for Computational Linguistics.

Xiaodong Yu, Stephen Mayhew, Mark Sammons, and Dan Roth. 2018. On the strength of character language models for multilingual named entity recognition. *arXiv preprint arXiv:1809.05157*.

GuoDong Zhou and Jian Su. 2002. Named entity recognition using an hmm-based chunk tagger. In *proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 473–480. Association for Computational Linguistics.