Neural Entity Recognition with Gazetteer based Fusion

Anonymous ACL-IJCNLP submission

Abstract

Incorporating external knowledge into Named Entity Recognition (NER) systems has been widely studied in generic domain. In this paper, we focus on clinical domain where only limited data is accessible and interpretability is important. With recent advancement in technology and increased number of clinical trials has resulted in discovery of new drugs , procedures as well as medical conditions. These factors motivate towards building robust zero-shot NER systems. We propose an auxiliary gazetteer model and fuse it with NER system, which results in better robustness and interpretability across different clinical datasets. Our gazetteer based fusion model is data efficient, achieving +1.7 micro-F1 gains on i2b2 dataset using 20% training data, and brings + 4.7 micro-F1 gains on novel entity mentions never presented during training. Moreover, our fusion model is able to quickly adapt to new mentions in gazetteers without re-training and the gains from name knowledge are transferable to related datasets.

1 Introduction

Named entity recognition (NER) (Lample et al., 2016; Ma and Hovy, 2016) aims to identify text mentions of specific entity types. In clinical domain, it's particularly useful for automatic information extraction, e.g., diagnosis information and adverse drug events, which could be applied for a variety of downstream tasks such as clinical event surveillance, decision support (Jin et al., 2018), pharmacovigilance, and drug efficacy studies.

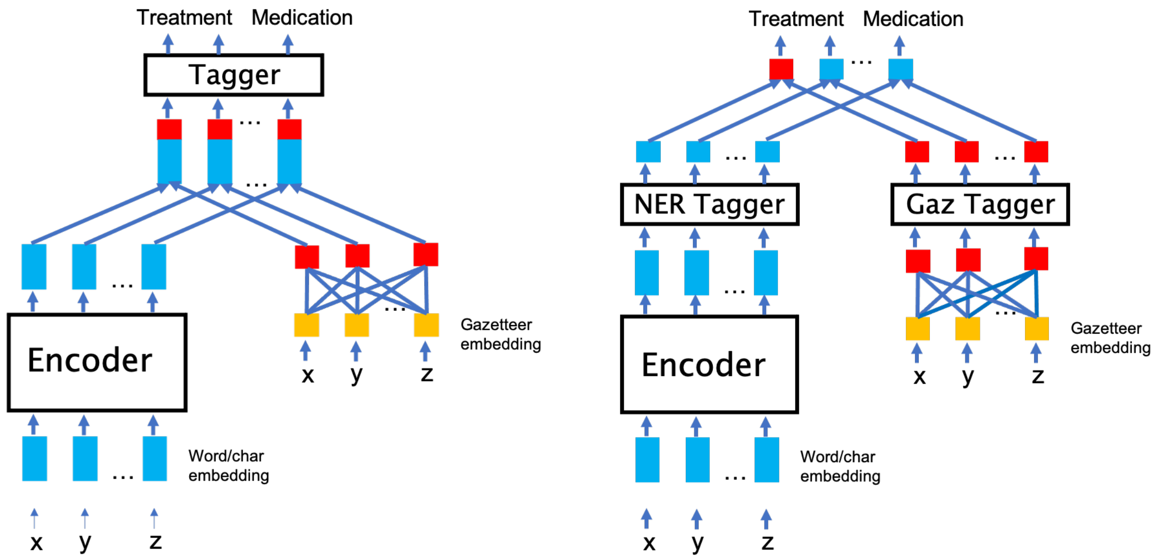
We have witnessed a rapid progress on NER models using deep neural networks. However, applying it to clinical domain raises several challenges: (a) only limited data is accessible, (b) new drugs, procedures and medical conditions are discovered and (c) building robust and explainable models is crucial. Motivated by this, we attempt to incorporate external name knowledge, e.g., *Remdesivir* is a `DRUG` and *COVID-19* is a `Medical Condition`, into neural NER models for clinical applications.

Recent work on leveraging external knowledge can be categorized into two categories - **Gazetteer embedding** and **Gazetteer model**. Song et al. (2020) feed the concatenation of BERT output and gazetteer embedding into Bi-LSTM-CRF. Peshterliev et al. (2020) use self-attention over gazetteer types to enhance gazetteer embedding and then concatenate it with ELMO, char CNN and Glove embeddings. **Gazetteer model.** The basic idea of **Gazetteer model** is to treat name knowledge as a new clinical modality. Magnolini et al. (2019) combine the output layers of BLSTM and gazetteer model and feed them into the CRF layer. Liu et al. (2019a) apply hybrid semi-Markov conditional random field (HSCRF) to predict a set of candidate spans and rescore them by pre-trained gazetteer model.

In this paper, we combine the advantages of both worlds. Unlike the work of Peshterliev et al. (2020), we build self-attention over entity mentions and their context rather than over different gazetteer types. For example, *Take Tylenol 3000* `(NUM)` *mg* `(METRIC)` *per day*, in which *Tylenol* is more likely to be a `DRUG` given `NUM`, `METRIC` in context. Moreover, we study two fusion methods to integrate information from two modalities.

- *Early fusion*. Similar to Magnolini et al. (2019), NER model and gazetteer model apply a shared tagger, as shown in Fig. 1a

- *Late fusion*. In order to provide better interpretability and flexibility, we allow NER model and gazetteer model to apply separate taggers and fuse information before taking softmax, as shown in Fig. 1b



(a) Early fusion

(b) Late fusion

Figure 1: Model Architecture. (a) Early fusion. The outputs of NER and gazetteer are concatenated and fed into a shared tagger. (b) Late fusion. NER and gazetteer apply separate taggers and two modalities are fused by taking element-wise max pooling.

Unlike the work of Liu et al. (2019a), NER model and gazetteer model are jointly learned end-to-end.

Our contributions are as follows. (1) We propose to argument NER model with an auxiliary gazetteer model via *late fusion*, which provides better interpretability and flexibility. Interestingly, the NER model can preserve the gains even if the gazetteer model is unplugged at test time. (2) A thorough of analysis shows that fusion model is data efficient, explainable and is able to quickly adapt to new entity mentions in gazetteers. (3) Experiments show that fusion model consistently brings gains cross different clinical NER datasets.

2 Approach

2.1 NER model

NER is a sequence tagging problem by maximizing a conditional probability of tags $\mathbf{y}$ given an input sequence $\mathbf{x}$. We first encode $\mathbf{x}$ into hidden vectors and apply a tagger to produce output $\mathbf{y}$.

$$\mathbf{r} = \text{Encoder}_R\big(\mathbf{x}\big) \quad (1)$$
$$\mathbf{o}_t^r = \text{Tagger}_R\big(\mathbf{r}_t\big) \quad (2)$$
$$\mathbf{y}_t = softmax(\mathbf{o}_t^r) \quad (3)$$

2.2 Gazetteer model

We embed gazetteers into $\mathbf{E} \in \mathbb{R}^{M \times K \times d}$, where M is the number of gazetteers, K is the number of labels, and d is the embedding size. In order to model the association of name knowledge between entity mentions and their contexts, we compute context-aware gazetteer embedding $\mathbf{g}_t$, for token $\mathbf{x}_t$, using scaled dot-product self-attention

$$\mathbf{g}_t = \text{softmax}\Big(\frac{\mathbf{E}_t^g(\mathbf{E}_{t'}^g)^T}{\sqrt{d}}\Big)\mathbf{E}_t^g,\ \forall |t - t'| \leq w \quad (4)$$

where $\mathbf{E}_t^g = [\mathbf{E}_{0,z_t^0}; \mathbf{E}_{1,z_t^1}; \cdots, \mathbf{E}_{M,z_t^M}]$, z_t^j is the gazetteer label of token $\mathbf{x}_t$ in gazetteer j, and w is the size of attention window.

Similar to NER model, we apply a tagger to produce output $\mathbf{y}$

$$\mathbf{o}_t^g = \text{Tagger}_G\big(\mathbf{g}_t\big) \quad (5)$$
$$\mathbf{y}_t = softmax(\mathbf{o}_t^g) \quad (6)$$

2.3 Fusion: NER + gazetteer

To better use information from both modalities, we investigate two different fusion methods to combine information from NER and gazetteer.

- *Early fusion.* In Fig.1a, we concatenate $\mathbf{r}_t$ with $\mathbf{g}_t$, and feed it into a shared tagger

$$\mathbf{y}_t = softmax\Big(\text{Tagger}_{RG}\big([\mathbf{r}_t; \mathbf{g}_t]\big)\Big) \quad (7)$$

- *Late fusion.* In Fig.1b, we directly fuse $\mathbf{o}_t^r$ and $\mathbf{o}_t^g$ by performing element-wise max pooling

$$\mathbf{y}_t = softmax\big(\max(\mathbf{o}_t^r, \mathbf{o}_t^g)\big) \quad (8)$$

3 Experiments

3.1 Experimental setup

LM pre-training. We continue to pre-train RoBERTa$_{\text{base}}$ (L=12, H=768, A=12) (Liu et al.,

Table 1: Results on i2b2 (Med, TTP) and DCN (Med, DS). We report micro-F1 score, each is averaged over 3 random seeds.

	i2b2		DCN	
	Med	TTP	Med	DS
NER w/o fusion	92.26	87.22	84.51	83.99
Early fusion	92.14	87.42	84.82	84.51
Early fusion + attention	92.44	87.43	84.99	84.47
Late fusion	92.37	87.32	84.84	84.58
Late fusion + attention	92.35	87.41	84.82	84.37

2019b) on **M**edical **I**nformation **M**art for **I**ntensive **C**are (MIMIC-III) dataset (Johnson et al., 2016), which comprises deidentified clinical data from approximately 60k intensive care unit admissions.

Fine-tuning on clinical NER datasets. We fine-tune $\text{RoBERTa}_{\text{mimic}}$ and learn gazetteer model (w/ NER tagger) from scratch on clinical datasets.

- i2b2 - We use public datasets from the 2009 and 2010 i2b2 challenges for medication (Med) (Uzuner et al., 2010), and "test, treatment, problem" (TTP) entity extraction. We follow the original data split from (Chalapathy et al., 2016) of 170 notes for training and 256 for testing.
- De-identified clinical notes (DCN) - Second dataset (Bhatia et al., 2018) consists of 1,500 de-identified, annotated clinical notes with medications (Med) and medical conditions (DS). We follow i2b2 challenge guidelines for data annotation.

We employ a tagging scheme that follows an inside, outside, begin, end and singleton (IOBES) format. We extract gazetteer dictionaries: medical condition and drugs from xx.

We minimize the cross-entropy loss during training and report micro-F1 score at test time. We use $\text{RoBERTa}_{\text{mimic}}$ as NER encoder and parameterize Taggers via Multi-layer Perception (MLPs). We use BertAdam optimizer, learning rate 0.00005, and dropout 0.1. We tune hyper-parameters $d \in [2, 12]$ and $w \in [2, 10]$ on validation set.

3.2 Results.

We report overall results in Table 1. We observe that incorporating name knowledge consistently boost performance on all datasets by $0.18 \sim 0.59$ micro-F1 gains. Overall, two fusion methods achieve comparable results.

Table 2: Cross-evaluation on i2b2 Med and DCN Med. Column: dataset models are trained on. Row: dataset models are evaluated on.

	i2b2	DCN
i2b2	94.54 → 94.77 (+0.23)	68.78 → 69.68 (+0.9)
DCN	59.98 → 60.08 (+0.1)	90.02 → 90.71 (+0.69)

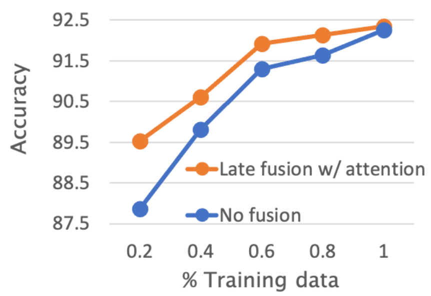


Figure 2: Accuracy vs. Training data size on i2b2 Med. We randomly sample 20%, 40%, $\cdots$, 100% of training data and report micro-F1 score averaged over 3 random seeds.

3.3 Analysis

We investigate the effectiveness of late fusion on handling three challenges: little data, novel entity mentions and interpretability.

3.3.1 Limited data access

Typically, only very limited data is ready to use in clinical domain. In this section, we focus on evaluating fusion model in low-resource settings as well as investigate whether the gain is transferable across related datasets. Here we present results with late fusion methodology.

Low-resource setting - We evaluate late fusion by reducing training data size from 100% to 20%. Fig.2 shows late fusion gains more when less training data is present. With 20% training data, late fusion is able to boost performance over the baseline model by 1.7 micro-F1 on i2b2 medication dataset.

Transfer learning. To verify the generalization ability of late fusion, we train models on one dataset and report evaluation on another data source. We re-train models on i2b2 Med and DCN Med only using common entity types: Dosage, Medication, Frequency, and Mode. Table.2 shows that the gains from gazetteer enhanced fusion models are preserved in i2b2 → DCN and DCN → i2b2.

Table 3: Performance on unseen entity mentions. Models are trained using 20% training data. We report performance of Medication entity in i2b2 Med and Treatment in i2b2 TTP.

	Medication	Treatment
NER w/o fusion	76.96	72.40
Late fusion w/ attention	**81.63 (+4.7)**	**74.30 (+1.9)**

Table 4: Ablation study on individual modules.

R_0	RG	R	R_0G
76.23	96.33	90.71	85.75

3.3.2 Novel entity mentions

New drugs and medical condition come out very frequently. For example, "remdesivir"and "Baricitinib" for COVID-19. To investigate the effect of late fusion on unseen entity mentions, we focus on answering questions: whether it can generalize well on unseen entity mentions? and whether it is able to correct prediction once novel entity names are added into gazetteer without re-training?

Zero-shot. We report results on unseen entity mentions that never presented in train and validation sets. In Table 3, we see that late fusion brings significant improvement: +4.7 F1 for medication entity (i2b2 Med) and +1.9 F1 for Treatment entity (i2b2 TTP). These results are useful with accelerated growth in drug development as well as in practical settings where entity extraction is one of the components to build knowledge graph and search engines.

"One"-shot in gazetteer. We evaluate the ability of late fusion to quickly adapt to non-stationary gazetteers, e.g., specialists might add new entity mentions into gazetteers or give feedback when models make incorrect prediction.

For this analysis, we split entity mentions in training set into two parts: 70% labelled and 30% in gazetteer, and compare models:

- $\mathbf{R}_0$: NER model only
- $\mathbf{RG}$: $\mathbf{R}$ + $\mathbf{G}$ via late fusion
- $\mathbf{R}$: Unplug $\mathbf{G}$ from $\mathbf{RG}$ after training
- $\mathbf{R}_0\mathbf{G}$: Fix $\mathbf{R}_0$ and learn $\mathbf{G}$ via late fusion

where $\mathbf{R}$ is NER model and $\mathbf{G}$ is gazetteer model. In Table 4, we observe that

- $\mathbf{R} > \mathbf{R}_0$. $\mathbf{G}$ can regularize $\mathbf{R}$ during training.
- $\mathbf{R}_0\mathbf{G} > \mathbf{R}_0$ and $\mathbf{RG} > \mathbf{R}$. Besides serving as a regularizer, $\mathbf{G}$ brings gains at test time.

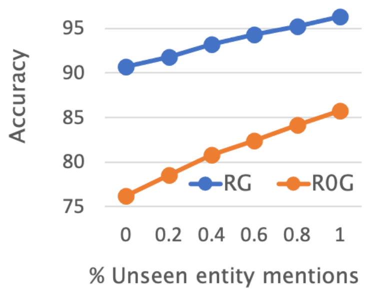


Figure 3: Quick adaptation to non-stationary gazetteers. As we increase the number of unseen entity mentions included in gazetteers, the performance goes up without re-training.

Table 5: Qualitative examples.

(1) Treated for $\underbrace{\text{COPD flare}}_{\text{B-R, I-R}}$ with supplemental DuoNebs	
NER w/o fusion	$\underbrace{\text{COPD flare}}_{\text{B-R, O}}$
Late fusion w/o attention	$\underbrace{\text{COPD flare}}_{\text{B-R, O}}$
Late fusion w/ attention	$\underbrace{\text{COPD flare}}_{\text{B-R, I-R}}$
(2) Postop day 0, increase $\underbrace{\text{sodium}}_{\text{S-M}}$, free water added	
$\mathbf{R}$: $\underbrace{\text{sodium}}_{\text{O}}$, G: $\underbrace{\text{sodium}}_{\text{S-M}}$, RG: $\underbrace{\text{sodium}}_{\text{S-M}}$	

Moreover, we evaluate late fusion by varying the number of unseen entity mentions included in gazetteers. In Fig.3, without re-training, we observe that late fusion can adapt to new mentions in gazetteers, almost linearly.

3.3.3 Interpretability

Explainable and controllable models are very important for clinical applications. Unfortunately, it is extremely challenging for deep neural networks. We illustrate two qualitative examples in Table.5. Late fusion models are trained on i2b2 Med using 20% training data.

(1) Late fusion correctly predicts *flare* as I-R (Reason) since *COPD flare* is Medical Condition.

(2) By looking into individual predictions from $\mathbf{R}$ and $\mathbf{G}$, we notice that correct prediction is caused by name knowledge in gazetteers.

Overall, late fusion provides us a tool to diagnosis system to answer questions whether NER or gazetteer model failed and explain why mentions belong to a particular entity type.

4 Conclusion

We studied fusion methods to improve NER system by leveraging name knowledge from gazetteers.

We did a through of analysis on the effectiveness of fusion methods on handling little data and non-stationary gazetteer dictionary. In addition, we demonstrated that fusion models are explainable and can be used to diagnosis NER system. Future research should extend our approach to structured knowledge to further improve NER system and gain better interpretability.

References

Parminder Bhatia, Busra Celikkaya, and Mohammed Khalilia. 2018. Joint entity extraction and assertion detection for clinical text. *arXiv preprint arXiv:1812.05270*.

Raghavendra Chalapathy, Ehsan Zare Borzeshi, and Massimo Piccardi. 2016. Bidirectional lstm-crf for clinical concept extraction. *arXiv preprint arXiv:1611.08373*.

Mengqi Jin, Mohammad Taha Bahadori, Aaron Colak, Parminder Bhatia, Busra Celikkaya, Ram Bhakta, Selvan Senthivel, Mohammed Khalilia, Daniel Navarro, Borui Zhang, et al. 2018. Improving hospital mortality prediction with medical named entities and multimodal learning. *arXiv preprint arXiv:1811.12276*.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

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.

Tianyu Liu, Jin-Ge Yao, and Chin-Yew Lin. 2019a. Towards improving neural named entity recognition with gazetteers. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5301–5307.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. *arXiv preprint arXiv:1603.01354*.

Simone Magnolini, Valerio Piccioni, Vevake Balaraman, Marco Guerini, and Bernardo Magnini. 2019. How to use gazetteers for entity recognition with neural models. In *Proceedings of the 5th Workshop on Semantic Deep Learning (SemDeep-5)*, pages 40–49.

Stanislav Peshterliev, Christophe Dupuy, and Imre Kiss. 2020. Self-attention gazetteer embeddings for named-entity recognition. *arXiv preprint arXiv:2004.04060*.

Chan Hee Song, Dawn Lawrie, Tim Finin, and James Mayfield. 2020. Improving neural named entity recognition with gazetteers. *arXiv preprint arXiv:2003.03072*.

Özlem Uzuner, Imre Solti, and Eithon Cadag. 2010. Extracting medication information from clinical text. *Journal of the American Medical Informatics Association*, 17(5):514–518.