BERT might be Overkill: A Tiny but Effective Biomedical Entity Linker based on Residual Convolutional Neural Networks

Tuan Lai, Heng Ji, ChengXiang Zhai
Department of Computer Science
University of Illinois Urbana-Champaign
{tuanml2, hengji, czhai}@illinois.edu

Abstract

Biomedical entity linking is the task of linking entity mentions in a biomedical document to referent entities in a knowledge base. Recently, many BERT-based models have been introduced for the task. While these models have achieved competitive results on many datasets, they are computationally expensive and contain about 110M parameters. Little is known about the factors contributing to their impressive performance and whether the over-parameterization is needed. In this work, we shed some light on the inner working mechanisms of these large BERT-based models. Through a set of probing experiments, we have found that the entity linking performance only changes slightly when the input word order is shuffled or when the attention scope is limited to a fixed window size. From these observations, we propose an efficient convolutional neural network with residual connections for biomedical entity linking. Because of the sparse connectivity and weight sharing properties, our model has a small number of parameters and is highly efficient. On five public datasets, our model achieves comparable or even better linking accuracy than the state-of-the-art BERT-based models while having about 60 times fewer parameters. [1]

1 Introduction

Biomedical entity linking (EL) (Zheng et al., 2014) is the task of linking biomedical mentions (e.g., diseases and drugs) to standard referent entities in a curated knowledge base (KB). For example, given the sentence "*The average* ***NH3*** *concentrations were low.*", the mention ***NH3*** should be linked to the entity `KB:Ammonia`. Biomedical EL is an important research problem, with applications in many downstream tasks, such as biomedical question answering (Lee et al., 2020), information retrieval, and information extraction (Wang et al., 2020; Huang et al., 2020; Lai et al., 2021b; Zhang et al., 2021). In general, two main challenges of the EL task are: (1) *ambiguity* - the same word or phrase can be used to refer to different entities; (2) *variety* - the same entity can be referred to by different words or phrases. Unlike in the general domain, mentions in biomedical documents are relatively unambiguous (D'Souza and Ng, 2015; Li et al., 2017). Building a system for biomedical EL involves primarily addressing the variety problem.

Recently, many BERT-based models have been introduced for biomedical EL (Ji et al., 2020; Sung et al., 2020; Liu et al., 2020, 2021). While these models can achieve state-of-the-art results on many biomedical EL datasets, they are computationally expensive and contain about 110M parameters. Even though there are scientific labs that have a lot of computing resources, many researchers still have minimal access to large-scale computational power (Strubell et al., 2019). Therefore, it is of practical importance to provide a more scalable solution for biomedical entity linking. Furthermore, the factors contributing to the success of these large BERT-based models remain unclear. And thus, it is not known whether the over-parameterization is needed to achieve competitive performance.

In this work, through a set of probing experiments, we shed some light on the inner workings of existing BERT models for biomedical EL. Surprisingly, the performance only changes slightly when the input word order is shuffled or when the attention scope is restricted. Based on these observations, we propose an effective convolutional neural network with residual connections (ResCNN) for the task. Because of the sparse connectivity and weight sharing properties, ResCNN has a small number of parameters and is highly efficient. Experiments on five datasets show that the performance of ResCNN is comparable to the state-of-the-art (SOTA) BERT-based models while having about

[1] The code is publicly available at `https://github.com/laituan245/rescnn_bioel`

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 1631–1639
November 7–11, 2021. ©2021 Association for Computational Linguistics

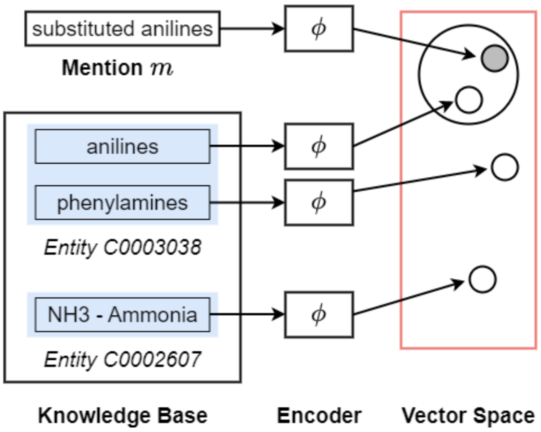

Figure 1: An illustration of the adopted approach to EL. In this example, the closest neighbor to the source mention is the entity name *anilines*. Therefore, this mention should be linked to the entity *C0003038*.

60 times fewer parameters.

2 Methods

In the following sections, we will first describe some preliminaries relating to the formulation of the EL problem and a general approach for the task (Sec. 2.1). We will then go into details about our probing experiments in Sec. 2.2. We will describe the design of our ResCNN model in Sec. 2.3.

2.1 Preliminaries

Problem Formulation Given an entity mention m from a biomedical text and a knowledge base (KB) consisting of N entities $\mathcal{E} = \{e_1, e_2, ..., e_N\}$, the task is to find the entity $e_i \in \mathcal{E}$ that m refers to. We assume that each entity is associated with a primary name and a list of alternative names. We denote the set of all names in the KB as $\mathcal{N} = \{n_1, n_2, ..., n_M\}$, where M is the number of names. We use T_m and T_{n_j} to denote the textual forms of m and n_j respectively. Except for a list of names for each entity, we do not assume the availability of any other information in the KB (e.g., entity types or description sentences). Our formulation is general and suitable for a wide range of real-world settings.

General Approach A general approach to EL is to train an encoder ϕ that encodes mentions and entity names into the same vector space (Gillick et al., 2019) (Figure 1). Before inference, we use ϕ to pre-compute embeddings for all the entity names in the KB. During inference, mentions are also encoded by ϕ and entities are retrieved using a simple distance function such as cosine similarity. In this work, we adopt this general approach, because it is more efficient and simpler than the two-stage retrieval and re-ranking systems (Wu et al., 2020). Several recent SOTA methods for biomedical EL also follow this approach. For example, Liu et al. (2020) models ϕ using SAPBERT, a BERT model pretrained on UMLS synonyms:

$$\begin{aligned} \phi(m) &= \text{SAPBERT}_{\text{CLS}}(T_m) \\ \phi(n_j) &= \text{SAPBERT}_{\text{CLS}}(T_{n_j}) \; \forall n_j \in \mathcal{N} \end{aligned} \quad (1)$$

where $\text{SAPBERT}_{\text{CLS}}$ returns the final hidden state corresponding to the [CLS] token. Since SAPBERT was pre-trained on almost 12M pairs of synonyms, it can be directly used without further fine-tuning on the target task's training data. However, for several datasets, the performance can still be improved by training with task-specific supervision.

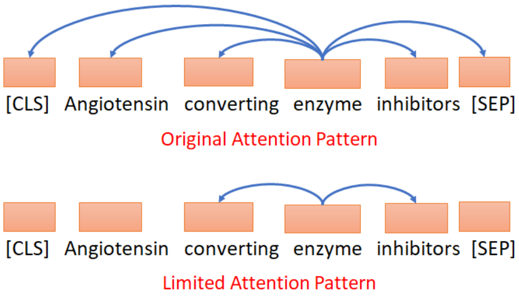
[CLS] Angiotensin converting enzyme inhibitors [SEP]
Original Attention Pattern
[CLS] Angiotensin converting enzyme inhibitors [SEP]
Limited Attention Pattern

Figure 2: Attention scope restriction. In this example, the window size of the limited attention head is 3.

2.2 Probing Experiments

Previous studies have shown that BERT can encode a wide range of syntactic and semantic features (Tenney et al., 2019; Jawahar et al., 2019). However, it is unknown to what extent existing BERT models for biomedical EL utilize such rich linguistic signals. We take the first step towards answering this question by investigating the most basic aspects.

Word Order Permutation We analyze whether BERT models fine-tuned for biomedical EL even consider one of the most fundamental properties of a sequence - the word order. In this probing experiment, we first train an EL model on the original (unshuffled) training set of a dataset. We then evaluate the model on the development set under the condition that the tokens of each mention/entity-name are shuffled.

Attention Scope Restriction The self-attention mechanism of BERT makes each token in the input directly interact with every other token (Vaswani et al., 2017). As a result, the attention operation is quadratic to the input length. To analyze whether direct connections between distant tokens are crucial for biomedical EL, we conduct experiments where we restrict the attention scope to a local window (Figure 2). We first train a BERT-based EL model on the provided training set of a dataset. We use the original attention mechanism during training. During evaluation, we limit the attention scope to a fixed window size by applying a masking operation:

$$\mathbf{M}[i,j] = \begin{cases} 1, & \text{if } |i-j| \leq \lfloor w/2 \rfloor \\ \text{-}\infty, & \text{otherwise} \end{cases}$$
$$\text{Attention}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \text{softmax}\left(\frac{\mathbf{M}\odot\mathbf{QK}^T}{\sqrt{p}}\right)\mathbf{V} \tag{2}$$

where **Q**, **K**, and **V** are the matrices of the queries, keys, and values (respectively) of an attention head (Vaswani et al., 2017). w denotes the window size (w is odd), $\odot$ denotes element-wise multiplication, and p is a scaling factor. We restrict the attention scope of every token at every layer except for the [CLS] token at the last layer. We let the token attend to every other token at the last layer.

2.3 ResCNN for Biomedical Entity Linking

As to be discussed in Section 3, the performance of existing BERT models only changes slightly when the input word order is shuffled or when the attention scope is limited. These observations suggest that a simpler model that mainly focuses on capturing local interactions may perform as well as SOTA BERT-based models. A natural candidate that exhibits the desired properties is the convolutional neural network (CNN) architecture. CNNs have been empirically shown to be quite effective in capturing local features (Kim, 2014). Furthermore, CNNs typically use fewer parameters than Transformer-based models because of their sparse connectivity and weight sharing properties. To this end, we introduce a simple but effective CNN with residual connections (ResCNN) for biomedical EL. Given an input text (e.g., a query mention or an entity name), ResCNN computes a vector representation for the input through several layers.

Token Embedding Layer We first use the BERT WordPiece tokenizer (Wu et al., 2016) to split the original input text into a sequence of tokens. We then transform each token into an initial vector representation by re-using the first embedding layer of PubMedBERT (Gu et al., 2020). This operation is very similar to using traditional word embeddings such as GloVe (Pennington et al., 2014), and so it can be carried out efficiently. We keep the embedding layer fixed and do not tune its parameters during training. An advantage of WordPiece tokenization is that a relatively small vocabulary (e.g., 30,000 wordpieces) is sufficient to model large, naturally-occurring corpora. In contrast, the vocabulary size of traditional word embeddings is typically much larger.

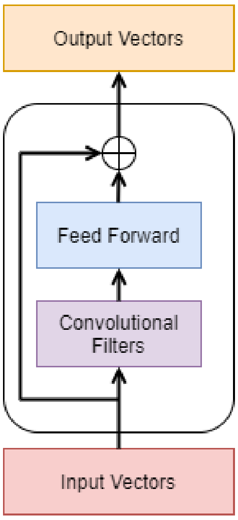

Figure 3: Encoding block of ResCNN.

Encoding Layer Our encoding layer consists of several *encoding blocks* (Figure 3). Each block has multiple convolutional filters of varying window sizes (Kim, 2014). Each filter is followed by an ReLU activation. We also employ a position-wise fully connected feed-forward network after applying the convolutional filters. In addition, there is a residual connection between the input and output of each encoding block. Residual connections alleviate the vanishing gradient problem (He et al., 2016). Overall, our encoding blocks are quite similar to the Transformer encoder layers (Vaswani et al., 2017). However, we use local convolutional filters for feature extraction instead of the global attention mechanism.

Pooling Layer To obtain the final vector representation for the input, we apply a pooling operation. In this work, we experiment with two different pooling strategies: (1) *Max Pooling* (Kim, 2014) (2) *Self-Attention Pooling* (Zhu et al., 2018).

Models	Top-1 Accuracy (on development sets)					Avg. % change
	NCBI-d	BC5CDR-d	BC5CDR-c	MedMentions	COMETA	
SAPBERT (Fine-Tuned) (2020)	91.1	90.9	98.2	54.4	74.9	
Word Order Permutation						
◆ Shuffle unigrams	88.2	90.2	94.0	53.2	65.6	-4.58%
◆ Shuffle bigrams	89.1	90.8	96.4	53.8	71.9	-1.87%
◆ Shuffle trigrams	90.5	91.0	97.7	54.0	73.1	-0.87%
Attention Scope Restriction						
■ Context size = 3	91.1	90.3	97.9	53.2	71.9	-1.44%
■ Context size = 5	91.2	90.9	97.6	53.8	73.4	-0.74%

Table 1: Results of our conducted probing experiments with SAPBERT (Liu et al., 2020).

Models	Top-1 Accuracy (on test sets)			Avg. % change
	NCBI-d	BC5CDR-d	BC5CDR-c	
BIOSYN (Dense) (Sung et al., 2020)	90.7	92.9	96.6	
Word Order Permutation				
◆ Shuffle unigrams	67.0	77.0	74.8	-21.94%
◆ Shuffle bigrams	77.7	87.2	85.6	-10.62%
◆ Shuffle trigrams	82.7	91.4	92.2	-5.0%
Attention Scope Restriction				
■ Context size = 3	81.0	84.5	96.5	-6.61%
■ Context size = 5	78.8	87.5	96.5	-6.35%

Table 2: Results of our conducted probing experiments with BIOSYN (Sung et al., 2020).

We acknowledge that most of the components of our model are not novel as CNNs with residual links have been used in other tasks (Conneau et al., 2017; Huang and Wang, 2017). Nevertheless, our work provides evidence for the importance of carefully justifying the complexity of existing or newly proposed models. Depending on the specific task, a lightweight model may perform as well as the large BERT-based models. Also, our proposed ResCNN achieves SOTA performance on several datasets while being even more efficient than previous CNN-based or RNN-based methods (Sec. 3).

3 Experiments

Data and Experimental Setup We experiment across five different datasets: NCBI (Dogan et al., 2014), BC5CDR-c and BC5CDR-d (Li et al., 2016), MedMentions (Mohan and Li, 2019), and COMETA (Basaldella et al., 2020). For each dataset, we follow the data split by Liu et al. (2020). It is worth highlighting that even though the five datasets can all be categorized as "biomedical datasets", they have very different characteristics. For example, while MedMentions was constructed by annotating scientific papers, COMETA was built by crawling Reddit (a social media forum). We report results in terms of top-1 accuracy. Details about the hyperparameters are in the appendix.

Probing Results (SAPBERT) Table 1 shows the results of our probing experiments with SAPBERT (Liu et al., 2020). When the inputs' unigrams are randomly re-ordered, the performance of SAPBERT only drops by about 4.58% on average. The difference is even less noticeable when we shuffle trigrams instead of unigrams. Therefore, SAPBERT is highly insensitive to word-order randomization. These results agree with recent studies on general-domain BERT models (Pham et al., 2020; Sinha et al., 2021). Table 1 also shows that the performance of SAPBERT only changes slightly when the attention scope is limited.

Probing Results (BIOSYN) We have also experimented with BERT models trained on the BIOSYN framework (Sung et al., 2020). We directly use the trained BERT models downloaded from `https://github.com/dmis-lab/BioSyn`. Table 2 shows the results of our conducted probing experiments with BIOSYN. Note that the authors of BIOSYN only provided the trained checkpoints for NCBI-d, BC5CDR-d, and BC5CDR-c. Overall, the changes are more prominent for models trained on BIOSYN than for SAPBERT. Nevertheless, the performance only drops by about 5.0% on average

Models	Top-1 Accuracy (on test sets)					Nb. Parameters
	NCBI-d	BC5CDR-d	BC5CDR-c	MedMentions	COMETA	
BNE (2019)	87.7	90.6	95.8	-	-	4.1M
CNN-based Ranking (2020)	89.6	-	-	-	-	4.6M
SAPBERT (Fine-Tuned) * (2020)	92.3	93.2	96.5	50.4	75.1	110M
BIOSYN * (2020)	91.1	93.2	96.6	OOM	71.3	110M
BIOSYN (init. w/ SAPBERT) *	**92.5**	**93.6**	96.8	OOM	77.0	110M
ResCNN (Self-Attention Pooling)	92.2	93.2	**96.9**	**55.0**	79.4	**1.8M**
ResCNN (Max Pooling)	92.4	93.1	96.8	53.5	**80.1**	**1.7M**

Table 3: Overall test results on the five biomedical EL datasets. "-" denotes results not reported in the cited paper. The symbol * denotes BERT-based models. OOM stands for out-of-memory.

Models	NCBI-d		BC5CDR-d		BC5CDR-c		MedMentions		COMETA	
	CPU	GPU	CPU	GPU	CPU	GPU	CPU	GPU	CPU	GPU
SAPBERT (2020)	534s	58s	551s	66s	3478s	276s	OOM	6269s	6156s	470s
ResCNN + Max Pooling	33s	18s	35s	21s	169s	69s	3274s	1565s	289s	109s
Speedup (compared to SAPBERT)	16.2x	3.2x	15.7x	3.1x	20.6x	4.0x	-	4.0x	21.3x	4.3x

Table 4: Inference time of different models on CPU and GPU. OOM stands for out-of-memory.

when the inputs' trigrams are randomly re-ordered. The performance also only changes by 6.61% when the attention window is set to be 3.

Entity Linking Accuracy Table 3 shows the linking performance of various models. Despite having less than 2M parameters, our CNN-based models achieve better results than the previous BERT-based SOTA systems on three of the datasets. It is worth noting that SAPBERT (Liu et al., 2020) was pre-trained on almost 12M pairs of UMLS synonyms. Without such pre-training, our lightweight models still match the performance of SAPBERT.

Inference Time Table 4 shows the speed of various models on CPU and on GPU. Compared to SAPBERT, our model is about 3 to 4 times faster on GPU and about 15 to 20 times faster on CPU. It takes less time to run our model on CPU than running SAPBERT on GPU. These results demonstrate the efficiency of our proposed model.

4 Conclusions and Future Work

Our work has shown that while BERT has been widely used for many NLP tasks, it is sometimes an overkill for some tasks, in which case, a simpler model can be as effective as BERT and is often much more efficient. An interesting future direction is to study further how to systematically simplify/compress BERT based on the insights obtained using probing experiments to increase efficiency while maintaining effectiveness. We plan to extend our work to other domains as well as other information extraction tasks (Lai et al., 2020; Lin et al., 2020; Wen et al., 2021; Lai et al., 2021a; Li et al., 2020).

Acknowledgement

This research is based upon work supported by the Molecule Maker Lab Institute: An AI Research Institutes program supported by NSF under Award No. 2019897, NSF No. 2034562, Agriculture and Food Research Initiative (AFRI) grant no. 2020-67021-32799/project accession no.1024178 from the USDA National Institute of Food and Agriculture, and U.S. DARPA KAIROS Program No. FA8750-19-2-1004. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of DARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

Marco Basaldella, Fangyu Liu, Ehsan Shareghi, and Nigel Collier. 2020. COMETA: A corpus for medical entity linking in the social media. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3122–3137, Online. Association for Computational Linguistics.

Lihu Chen, G. Varoquaux, and Fabian M. Suchanek. 2020. A lightweight neural model for biomedical entity linking. *ArXiv*, abs/2012.08844.

Alexis Conneau, Holger Schwenk, Loïc Barrault, and Yann Lecun. 2017. Very deep convolutional networks for text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1107–1116, Valencia, Spain. Association for Computational Linguistics.

Rezarta Islamaj Dogan, Robert Leaman, and Z. Lu. 2014. Ncbi disease corpus: A resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 47:1–10.

Jennifer D'Souza and Vincent Ng. 2015. Sieve-based entity linking for the biomedical domain. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 297–302, Beijing, China. Association for Computational Linguistics.

Daniel Gillick, Sayali Kulkarni, Larry Lansing, Alessandro Presta, Jason Baldridge, Eugene Ie, and Diego Garcia-Olano. 2019. Learning dense representations for entity retrieval. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 528–537, Hong Kong, China. Association for Computational Linguistics.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2020. Domain-specific language model pretraining for biomedical natural language processing. *ArXiv*, abs/2007.15779.

C. Harris, K. J. Millman, S. Walt, Ralf Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, Nathaniel J. Smith, R. Kern, Matti Picus, S. Hoyer, M. Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fern'andez del R'io, Mark Wiebe, P. Peterson, Pierre G'erard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, Christoph Gohlke, and T. E. Oliphant. 2020. Array programming with numpy. *Nature*, 585 7825:357–362.

Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Kung-Hsiang Huang, Mu Yang, and Nanyun Peng. 2020. Biomedical event extraction with hierarchical knowledge graphs. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1277–1285, Online. Association for Computational Linguistics.

Yi Yao Huang and William Yang Wang. 2017. Deep residual learning for weakly-supervised relation extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1803–1807, Copenhagen, Denmark. Association for Computational Linguistics.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Zongcheng Ji, Qiang Wei, and H. Xu. 2020. Bert-based ranking for biomedical entity normalization. *AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science*, 2020:269–277.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Tuan Lai, Heng Ji, Trung Bui, Quan Hung Tran, Franck Dernoncourt, and Walter Chang. 2021a. A context-dependent gated module for incorporating symbolic semantics into event coreference resolution. *arXiv preprint arXiv:2104.01697*.

Tuan Lai, Heng Ji, ChengXiang Zhai, and Quan Hung Tran. 2021b. Joint biomedical entity and relation extraction with knowledge-enhanced collective inference. *arXiv preprint arXiv:2105.13456*.

Tuan Manh Lai, Trung Bui, Doo Soon Kim, and Quan Hung Tran. 2020. A joint learning approach based on self-distillation for keyphrase extraction from scientific documents. *arXiv preprint arXiv:2010.11980*.

Jinhyuk Lee, Sean S. Yi, Minbyul Jeong, Mujeen Sung, WonJin Yoon, Yonghwa Choi, Miyoung Ko, and Jaewoo Kang. 2020. Answering questions on COVID-19 in real-time. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online. Association for Computational Linguistics.

Haodi Li, Q. Chen, Buzhou Tang, Xiaolong Wang, H. Xu, Baohua Wang, and Dong Huang. 2017. Cnn-based ranking for biomedical entity normalization. *BMC Bioinformatics*, 18.

J. Li, Yueping Sun, Robin J. Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, A. P. Davis, C. Mattingly, Thomas C. Wiegers, and Z. Lu. 2016. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database: The Journal of Biological Databases and Curation*, 2016.

Manling Li, Ying Lin, Tuan Manh Lai, Xiaoman Pan, Haoyang Wen, Sha Li, Zhenhailong Wang, Pengfei Yu, Lifu Huang, Di Lu, et al. 2020. Gaia at smkbp 2020-a dockerlized multi-media multi-lingual knowledge extraction, clustering, temporal tracking and hypothesis generation system. In *Proceedings of Thirteenth Text Analysis Conference (TAC 2020)*.

Ying Lin, Heng Ji, Fei Huang, and Lingfei Wu. 2020. A joint neural model for information extraction with global features. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7999–8009, Online. Association for Computational Linguistics.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2020. Self-alignment pre-training for biomedical entity representations. *ArXiv*, abs/2010.11784.

Fangyu Liu, Ivan Vulić, Anna Korhonen, and Nigel Collier. 2021. Learning domain-specialised representations for cross-lingual biomedical entity linking. *arXiv preprint arXiv:2105.14398*.

S. Mohan and Donghui Li. 2019. Medmentions: A large biomedical corpus annotated with umls concepts. *ArXiv*, abs/1902.09476.

Adam Paszke, S. Gross, Francisco Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, Alban Desmaison, Andreas Köpf, E. Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, B. Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*.

Jeffrey Pennington, R. Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*.

Thang M. Pham, Trung Bui, Long Mai, and Anh M Nguyen. 2020. Out of order: How important is the sequential order of words in a sentence in natural language understanding tasks? *ArXiv*, abs/2012.15180.

Minh C. Phan, Aixin Sun, and Yi Tay. 2019. Robust representation learning of biomedical names. In *ACL*.

Koustuv Sinha, Robin Jia, Dieuwke Hupkes, Joelle Pineau, Adina Williams, and Douwe Kiela. 2021. Masked language modeling and the distributional hypothesis: Order word matters pre-training for little. *arXiv preprint arXiv:2104.06644*.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.

Mujeen Sung, Hwisang Jeon, Jinhyuk Lee, and Jaewoo Kang. 2020. Biomedical entity representations with synonym marginalization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3641–3650, Online. Association for Computational Linguistics.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *ArXiv*, abs/1706.03762.

Qingyun Wang, Manling Li, X. Wang, Nikolaus Nova Parulian, Guangxing Han, Jiawei Ma, Jingxuan Tu, Ying Lin, H. Zhang, Weili Liu, Aabhas Chauhan, Yingjun Guan, Bangzheng Li, Ruisong Li, Xiangchen Song, Huai zhong Ji, Jiawei Han, Shih-Fu Chang, J. Pustejovsky, D. Liem, A. El-Sayed, Martha Palmer, Jasmine Rah, C. Schneider, and B. Onyshkevych. 2020. Covid-19 literature knowledge graph construction and drug repurposing report generation. *ArXiv*, abs/2007.00576.

Haoyang Wen, Ying Lin, Tuan Lai, Xiaoman Pan, Sha Li, Xudong Lin, Ben Zhou, Manling Li, Haoyu Wang, Hongming Zhang, et al. 2021. Resin: A dockerized schema-guided cross-document cross-lingual cross-media information extraction and event tracking system. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, pages 133–143.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2020. Scalable zero-shot entity linking with dense entity retrieval. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6397–6407, Online. Association for Computational Linguistics.

Y. Wu, M. Schuster, Z. Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, M. Krikun, Yuan Cao, Q. Gao, Klaus Macherey, J. Klingner, Apurva Shah, M. Johnson, X. Liu, Lukasz Kaiser, Stephan Gouws, Y. Kato, Taku Kudo, H. Kazawa, K. Stevens, George Kurian, Nishant Patil, W. Wang, C. Young, J. Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, G. Corrado, Macduff Hughes, and J. Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *ArXiv*, abs/1609.08144.

Zixuan Zhang, Nikolaus Parulian, Heng Ji, Ahmed Elsayed, Skatje Myers, and Martha Palmer. 2021. Fine-grained information extraction from biomedical literature based on knowledge-enriched Abstract Meaning Representation. 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 6261–6270, Online. Association for Computational Linguistics.

Jin Guang Zheng, Daniel Howsmon, Boliang Zhang, Juergen Hahn, Deborah McGuinness, James Hendler, and Heng Ji. 2014. Entity linking for biomedical literature. In *BMC Medical Informatics and Decision Making*.

Yingke Zhu, Tom Ko, David Snyder, B. Mak, and Daniel Povey. 2018. Self-attentive speaker embeddings for text-independent speaker verification. In *INTERSPEECH*.

A Reproducibility Checklist

In this section, we present the reproducibility information of the paper.

Implementation Dependencies Libraries Pytorch 1.6.0 (Paszke et al., 2019), Transformers 4.4.2 (Wolf et al., 2020), Numpy 1.19.5 (Harris et al., 2020), CUDA 11.0.

Computing Infrastructure The experiments were conducted on a server with Intel(R) Xeon(R) Gold 5120 CPU @ 2.20GHz and NVIDIA Tesla V100 GPUs. The allocated RAM is 191.9G. GPU memory is 16G.

Datasets NCBI-d, BC5CDR-c, and BC5CDR-d can be downloaded from `https://github.com/dmis-lab/BioSyn`. MedMentions can be downloaded from `https://github.com/chanzuckerberg/MedMentions`. COMETA can be downloaded from `https://github.com/cambridgeltl/cometa`.

Average Runtime We have presented the information of the inference time of our models in the main paper.

Number of Model Parameters We have discussed about the models' sizes in the main paper.

Hyperparameters of Best-Performing Models Each of our best ResCNN models consists of 4 encoding blocks. Each encoding block has 100 filters of kernel size 1, 100 filters of kernel size 3, and 100 filters of kernel size 5 (300 filters in total). The learning rate used for training our models is set to be 0.001. We use the Adam optimizer to train the ResCNN models. We use Huggingface's Transformer library to experiment with different BERT models (Wolf et al., 2020).

Expected Validation Performance For each of the MedMentions and COMETA datasets, we report the test performance of the checkpoint with the best validation score in the main paper. For each of the remaining three datasets, we use the corresponding development (dev) set to search for the hyperparameters and then train on the train-dev (train+dev) set to report the final performance (Sung et al., 2020). The final validation scores of our ResCNN models are shown in the Table 5.

Models	Top-1 Accuracy (on test sets)					Nb. Parameters
	NCBI-d	BC5CDR-d	BC5CDR-c	MedMentions	COMETA	
ResCNN (Self-Attention Pooling)	92.9	97.0	99.5	55.0	79.3	1.8M
ResCNN (Max Pooling)	95.0	91.8	99.3	53.8	79.9	1.7M

Table 5: Final validation scores of our ResCNN models on the five biomedical EL datasets.