Benchmarking Biomedical Nested NER and Relation Extraction Models

Anonymous ACL submission

Abstract

The Open EPPI corpus comprises 151 full-text papers annotated by domain experts for entity mentions, protein-protein interactions (PPIs), and normalisation of entities to publicly available ontologies. The corpus is publicly available at [ANON]. We benchmark recent nested NER and relation extraction models. Results show that, although existing nested NER models achieve good performance on outermost and innermost entity mentions, they struggle with other types of nested mentions. Benchmark results for relation extraction show substantial room for improvement with precision under 70 and recall around 40 to 52.

1 Introduction

The increasing rate of biomedical publishing leads to a vicious cycle in which information overload can reduce the impact of the new knowledge (Aviv-Reuven and Rosenfeld, 2021). For example, within August 2021 alone, there were more than five hundred papers published at PubMed,[1] searching with the keyword "protein-protein interactions[Abstract]". Automatically extracting structured information—such as biomedical concepts, attributes, events, and their relations—from unstructured text can be a useful first step for researchers to find relevant information (Ananiadou and McNaught, 2006; Jiang, 2012).

Although impressive benchmark results have been observed in many Information Extraction (IE) datasets in the generic domain, such as ACE and OntoNotes corpora (Yamada et al., 2020; Wang et al., 2020; Zhong and Chen, 2021), reduced performance is usually reported when methods are applied to biomedical text (Leaman et al., 2015; Wei et al., 2016). There are several challenges associated with biomedical IE, due to the special subject matter of the content being discussed and the variety of language used in scholarly articles:

- Biomedical names may contain complex structure (e.g., nested) which cannot be easily recognised using standard NER models (Kim et al., 2003; Ju et al., 2018);
- Researchers tend to write long text to make the description more comprehensible and less confused, which requires the model to capture long-range contexts; and,
- The training of models usually requires sufficient amount of labelled data, which are usually difficult to obtain in the biomedical domain.

To facilitate ongoing research on biomedical IE, we introduce Open EPPI—a large-scale dataset for biomedical entity recognition, relation extraction and concept normalisation. The corpus contains a large portion of nested entity annotations—16.3% sentences contain nested mentions—and is therefore a suitable testbed for nested NER models. 24.7% PPIs in the corpus are inter-sentence relations, which requires the model to make use of long-range contexts to recognise them. The large size of the corpus—151 full text articles—also enables the development of more sophisticated document-level IE models. In the following sections, we describe first the corpus and then our benchmarking results using existing methods.

2 The Open EPPI corpus

The Open EPPI corpus contains 151 full-text articles under non-copyrighted license. The average number of sentences per article is 309.4, and the average number of words per sentence is 25.1. The units of annotation consist of entity mentions, relations and concept normalisations. The descriptive statistics are listed in Table 1.

The mentions annotated are either PROTEINS and other related entities involved in PPI relations, i.e., COMPLEX, FUSION, FRAGMENT

[1] https://pubmed.ncbi.nlm.nih.gov/

and MUTANT, or attributes of PPI relations, i.e., CELLLINE, DRUGCOMPOUND, EXPERIMENTALMETHOD, MODIFICATION. Annotators were able to nest mentions, however, mentions were not allowed to cross or to be discontinuous.[2] There are in total $22,609$ sentences containing at least one entity mention; $3,681$ (16.3%) sentences contain nested entity mentions. Note that there are also $2,332$ multi-type entity mentions (Dai, 2018). That is, one biomedical name may have multiple entity types. For example, one mention may be classified as both PROTEIN and DRUGCOMPOUND, indicating that the protein is used to affect the function of an organism, cell or biological process.

Two types of relations were annotated: PPI (interactions between two PROTEINS) and Frag (connect MUTANTS and FRAGMENTS with their parent PROTEINS). Annotators were permitted to mark relations between entities in the same sentence (intra-sentential) and in different sentences (inter-sentential). There are $2,796$ out of $11,309$ relations (24.7%) that link entity mentions located in different sentences; 713 even cross different paragraphs. Both positive and negative PPI relations, i.e., statements asserting that an interaction did or did not occur, were marked with properties used to distinguish between them.

Annotation was performed by nine biologists, all qualified to PhD level in biology. Fifty-one articles were annotated by two annotators, and twenty-five articles were annotated by three annotators. We refer readers to [ANON][3] for more details regarding document selection, annotation process and inter-annotator agreement analysis.

3 Evaluating NER models

We evaluate several existing NER models on the corpus: (1) the standard sequence tagging model which can handle only flat entity mentions; (2) a span-based nested NER model (Zhong and Chen, 2021); (3) a CRF-based nested NER model (Shibuya and Hovy, 2020); (4) a layered nested NER model (Wang et al., 2020); and, (5) a hypergraph based nested NER model (Wang and Lu, 2018).

	Train	Dev	Test
# Documents	114	15	22
# Sentences	35,520	4,724	6,468
# Tokens	892,836	116,043	164,353
# Mentions	51,247	7,058	8,765
# Relations	8,935	811	1,577

Table 1: The descriptive statistics of the corpus.

Since the standard sequence tagger cannot handle nested mentions directly, we follow the approach proposed by Ringland et al. (2019). That is, we train two flat NER models, using either the outermost (Flat Outermost in Table 2) or the innermost mentions (Flat Innermost in Table 2) for training. We also combine the outputs from these two flat NER models and denote the results as Flat Combined. Note that instead of the BiLSTM encoder used in (Ringland et al., 2019), we use SciBERT [4] (Beltagy et al., 2019) encoder due to its superiority. We refer readers to the aforementioned papers for more details of other nested NER models.

Evaluation metric We frame the task as a sentence-level NER task and use the mention-level micro F_1 score—requiring an exact match of mention start, end and entity type—as the main metric to evaluate the effectiveness of the model. The model checkpoint which is most effective on the development set, measured using the F_1 score, is used to evaluate the test set.

To evaluate the effectiveness of different models on recognising nested mentions, we also construct several subsets of the test set: (1) a subset where only sentences with nested mentions are included; (2) a subset where only multi-type entity mentions are considered; (3) a subset where only outermost mentions are considered; (4) a subset where only innermost mentions are considered; and, (5) a subset where only middle mentions are considered. To evaluate whether a model can recognised nested mentions simultaneously, we employ a new metric that calculates F_1 score over outer-inner mention pairs. For example, if there are four mentions: ABC (outermost), AB (contained by ABC and con-

[2] Discontinuous coordinations such as 'A and B cells' were annotated as two nesting entities 'A and B cells' and 'B cells'. This annotation strategy was also used in the GENIA corpus (Kim et al., 2003).

[3] In [ANON], lessons learned from annotating two corpora in the [ANON] project are discussed. The Open EPPI corpus is a non-copyrighted license subset of corpora annotated in the project.

[4] In our preliminary experiments, we find scibert-scivocab-uncased performs better than other pre-trained models, including those of larger size (e.g., bert-large-uncased). Therefore, we use this version of SciBERT in all experiments, except (Wang and Lu, 2018), which is based on GloVe embeddings and BiLSTM encoder.

	Single Mention						Mention pair
	All	Sent w. nest	Multi type	Outermost	Innermost	Middle	
	8,765	2,233	184	7,745	8,116	232	1,088
Flat Outermost	74.3	56.9	0.0	78.8	76.0	0.0	0.0
Flat Innermost	76.9	66.3	0.0	78.9	79.8	0.0	0.0
Flat Combined	77.5	70.2	28.8	79.9	80.0	25.6	25.6
(Wang and Lu, 2018)	58.5	59.2	62.3	59.0	58.9	56.0	36.3
(Wadden et al., 2019)	74.9	69.0	0.0	77.1	76.7	12.5	35.4
(Wang et al., 2020)	**78.9**	**77.4**	**67.7**	80.0	79.6	**61.5**	**54.0**
(Shibuya and Hovy, 2020)	78.8	76.9	65.5	**80.6**	79.8	60.7	45.4
(Zhong and Chen, 2021)	**78.9**	76.2	0.0	80.5	**80.3**	14.2	42.6
(Zhong and Chen, 2021) (LC)	**80.3**	76.4	0.0	**81.6**	**81.8**	2.5	45.4

Table 2: NER results on Open EPPI using different methods. The number of gold mentions belonging to each set and the number of nested mention pairs are listed in the table header. We frame the task as sentence-level NER, except in the LC (Larger Context) row, where sentences in the same paragraph are used to build contextual hidden representations.

taining B), B (innermost), and C (innermost), we want the model to recognise all outer-inner mention pairs: ABC-AB, ABC-B, ABC-C, AB-B, because these nested structure usually contain useful information (Ringland et al., 2019).

Main results [5] Table 2 shows the effectiveness of different NER models. First, we observe a decrease (ranging from 1.5 to 17.4) of effectiveness of all models—except (Wang and Lu, 2018)—when evaluated on subset of sentences containing nested mentions comparing to all sentences in the test set. Additionally, all models achieve low F_1 when evaluated on outer-inner mention pairs. This scenario demonstrates the difficulty of recognising nested mentions simultaneously. Secondly, sequence tagging based flat NER models achieve decent performance when evaluated on the complete test set. Flat Combined outperforms two nested NER models due to its superiority of recognising outermost and innermost entity mentions. The best performing model—(Zhong and Chen, 2021)—performs well on outermost and innermost mentions, but fails to deal with multi-type and middle layer mentions. Last but not least, several nested NER models achieve encouraging results on multi-type, which is comparatively less studied, and middle layer mentions.

Can larger context improve model effectiveness? Scholarly articles are usually well organised using the (Sub)Section-Paragraph-Sentence structure. There are on average 6.2 sentences forming a paragraph in the corpus. A natural question is whether framing the NER task on the paragraph level can improve the model effectiveness. That is, the model takes a paragraph—instead of a single sentence—as input and recognises all mentions within the paragraph. We find that this simplified setup [6] does not improve the model effectiveness. For example, the F_1 score of Flat Combined model decreases from 77.5 to 76.8; the F_1 score of (Wang et al., 2020) decrease slightly from 78.9 to 78.8. We also find more sophisticated way (Zhong and Chen, 2021) can benefit from larger context. That is, the task is still framed as sentence-level NER, recognising mentions within each sentence, however, the model can take as input other sentences in the same paragraph to build the hidden representations. This modification can bring moderate (1.4 F_1 score) improvements (Table 2), especially on outermost and innermost mentions.

Human performance Recall that there are 76 articles in the corpus which are annotated by at least two annotators. To estimate the task difficulty, we evaluate the performance of annotators using these multi-annotated articles. We consider one annotator's (the one who annotates most) annotations as ground truth and other annotators' annotations

[5] The evaluation scripts and predicted outputs from different models can be found at GitHub [ANON].

[6] There are 126 (2.3%) paragraphs contain more than 512 wordpieces. We truncate the first 512 wordpieces of these paragraphs to satisfy the limit of BERT maximum sequence length.

as ‘predicted outputs’, then we calculate F_1 score of these ‘predicted outputs’. The F_1 score evaluated on different documents range from 60.4 to 95.1, and the annotator-level F_1 score—averaging document-level scores by the same annotator—range from 79.8 to 89.4. We believe the annotator-level F_1 score can be considered as a performance target for NER models on the corpus.

4 Evaluating Relation Extraction Models

We consider two types of relation extraction settings: (1) classifying relations between two given mentions in the text; and (2) extracting triplets consisting of mentions and relations between them.

The former classification setting assumes the gold mentions are given. We use the performance under this simplified setting as a proxy for estimating upper bound of the performance of a more practical relation extraction model. Inspired by Wu and He (2019), we insert special tokens at both the beginning and end of two mentions, and simply use the output hidden states of SciBERT encoder corresponding to the first token in the sequence (i.e., [CLS]) as input of the final 3-way (Positive PPI, Negative PPI, Frag) classifier. This model achieves a very high accuracy of 96.8, which implies that scholarly articles are usually comprehensive enough to distinguish relations between described concepts.

Regarding the extraction setting, we evaluate (1) a pipeline approach (Zhong and Chen, 2021), where the relation classification model considers every pair of entities—the predicted outputs from a NER model—and predicts the relation type for each pair; and, (2) a joint approach (Wadden et al., 2019), where dynamic span graph is used to refine span representations, based on which NER and relation classification are performed. We apply *boundaries* evaluation (Bekoulis et al., 2018) where a predicted relation is correct if the boundaries of two predicted mentions and the predicted relation type are correct. Results in Table 3 show that there are still large improvement room for both methods, especially in terms of recall.

Model	P	R	F
(Wadden et al., 2019)	65.2	39.5	49.2
(Zhong and Chen, 2021)	67.8	51.5	58.5

Table 3: Relation extraction results on Open EPPI.

5 Related Work

Recent work has demonstrated increasing interest in nested entity recognition (Alex et al., 2007; Byrne, 2007; Finkel and Manning, 2009; Lu and Roth, 2015; Katiyar and Cardie, 2018; Fisher and Vlachos, 2019; Yu et al., 2020), however, corpora annotated with nested entity mentions are still rare. Besides, the most common data source of nested NER corpora is newswire (Mitchell et al., 2006; Walker et al., 2006; Benikova et al., 2014; Ringland et al., 2019; Plank et al., 2020). The well-known GENIA corpus (Kim et al., 2003) is a publicly available biomedical nested NER dataset. It consists of 2000 abstracts taken from MEDLINE database and contains around 100K entity annotations. The main difference between Open EPPI and GENIA is that entity mentions and relations between them are annotated on full-text articles in the former, whereas only abstracts are annotated with biological terms in the latter.

Many scientific IE datasets also share similar characteristics with Open EPPI, except that documents used in these datasets usually focus on different subject matters (Li et al., 2016; Augenstein et al., 2017; Gábor et al., 2018; Luan et al., 2018; Jain et al., 2020, 2021). CRAFT (Bada et al., 2012) consists of 67 full-text biomedical articles and contains concept mapping to different biomedical ontologies. It is worth noting that CRAFT contains discontinuous mentions—mentions composed of components that are separated by intervals—which present different challenges from the nested annotations in Open EPPI. Another difference between CRAFT and Open EPPI is that coreference annotations are provided in the former, whereas the latter considers semantic relations between different entity types. We believe the publicly released corpus, in combination with these existing corpora, enables a more comprehensive evaluation regarding challenges of different tasks and syntactic aspects.

6 Summary

We described Open EPPI, a biomedical information extraction corpus for nested NER and relation extraction. The corpus, evaluation scripts and benchmark results are publicly available at [ANON]. We hope they provide a valuable resource to develop tools and methods for automatically curating, reviewing and tracking the ever-increasing biomedical literature.

References

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

Sophia Ananiadou and John McNaught. 2006. *Text mining for biology and biomedicine*.

Isabelle Augenstein, Mrinal Das, Sebastian Riedel, Lakshmi Vikraman, and Andrew McCallum. 2017. SemEval 2017 Task 10: ScienceIE - Extracting Keyphrases and Relations from Scientific Publications. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 546–555, Vancouver, Canada.

Shir Aviv-Reuven and Ariel Rosenfeld. 2021. Publication patterns' changes due to the COVID-19 pandemic: a longitudinal and short-term scientometric analysis. *Scientometrics*, 126(8):6761–6784.

Michael Bada, Miriam Eckert, Donald Evans, Kristin Garcia, Krista Shipley, Dmitry Sitnikov, William A Baumgartner, K Bretonnel Cohen, Karin Verspoor, Judith A Blake, and Lawrence E Hunter. 2012. Concept annotation in the CRAFT corpus. *BMC Bioinformatics*, 13(1).

Giannis Bekoulis, Johannes Deleu, Thomas Demeester, and Chris Develder. 2018. Adversarial training for multi-context joint entity and relation extraction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2830–2836, Brussels, Belgium.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. 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 3615–3620, Hong Kong, China.

Darina Benikova, Chris Biemann, and Marc Reznicek. 2014. NoSta-D Named Entity Annotation for German: Guidelines and Dataset. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland.

Kate Byrne. 2007. Nested named entity recognition in historical archive text. In *International Conference on Semantic Computing (ICSC 2007)*, pages 589–596, Irvine, California.

Xiang Dai. 2018. Recognizing complex entity mentions: A review and future directions. In *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Student Research Workshop*, pages 37–44, Melbourne, Australia.

Jenny Rose Finkel and Christopher D Manning. 2009. Nested named entity recognition. In *Proceedings of the 2009 conference on empirical methods in natural language processing*, pages 141–150, Singapore.

Joseph Fisher and Andreas Vlachos. 2019. Merge and Label: A Novel Neural Network Architecture for Nested NER. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5840–5850, Florence, Italy.

Kata Gábor, Davide Buscaldi, Anne-Kathrin Schumann, Behrang QasemiZadeh, Haifa Zargayouna, and Thierry Charnois. 2018. SemEval-2018 Task 7: Semantic Relation Extraction and Classification in Scientific Papers. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, New Orleans, Louisiana.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven Q H Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P Lungren, Andrew Y Ng, Curtis P Langlotz, and Pranav Rajpurkar. 2021. RadGraph: Extracting Clinical Entities and Relations from Radiology Reports. In *The 35th Conference on Neural Information Processing Systems (NeurIPS 2021)*.

Sarthak Jain, Madeleine van Zuylen, Hannaneh Hajishirzi, and Iz Beltagy. 2020. SciREX: A Challenge Dataset for Document-Level Information Extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7506–7516, Online.

Jing Jiang. 2012. Information Extraction from Text. In *Mining Text Data*. Springer US.

Meizhi Ju, Makoto Miwa, and Sophia Ananiadou. 2018. A neural layered model for nested named entity recognition. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1446–1459, New Orleans, Louisiana.

Arzoo Katiyar and Claire Cardie. 2018. Nested named entity recognition revisited. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 861–871, New Orleans, Louisiana. NAACL.

J.-D. Kim, T Ohta, Y Tateisi, and J Tsujii. 2003. GENIA corpus–a semantically annotated corpus for biotextmining. *Bioinformatics*, 19(Suppl 1).

Robert Leaman, Ritu Khare, and Zhiyong Lu. 2015. Challenges in clinical natural language processing for automated disorder normalization. *Journal of biomedical informatics*, 57:28–37.

Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and

Zhiyong Lu. 2016. BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database : the journal of biological databases and curation*, 2016:baw068.

Wei Lu and Dan Roth. 2015. Joint mention extraction and classification with mention hypergraphs. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 857–867, Lisbon, Portugal.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-Task Identification of Entities, Relations, and Coreference for Scientific Knowledge Graph Construction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3219–3232, Brussels, Belgium.

Alexis Mitchell, Stephanie Strassel, Shudong Huang, and Ramez Zakhary. 2006. ACE 2004 Multilingual Training Corpus.

Barbara Plank, Kristian Nørgaard Jensen, and Rob van der Goot. 2020. DaN+: Danish Nested Named Entities and Lexical Normalization. In *Proceedings of the 28th International Conference on Computational Linguistics*, Stroudsburg, PA, USA.

Nicky Ringland, Xiang Dai, Ben Hachey, Sarvnaz Karimi, Cecile Paris, and James R Curran. 2019. NNE: A Dataset for Nested Named Entity Recognition in English Newswire. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5176–5181, Florence, Italy.

Takashi Shibuya and Eduard Hovy. 2020. Nested Named Entity Recognition via Second-best Sequence Learning and Decoding. *Transactions of the Association for Computational Linguistics*, 8:605–620.

David Wadden, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi. 2019. Entity, relation, and event extraction with contextualized span representations. 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 5784–5789, Hong Kong, China.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. ACE 2005 Multilingual Training Corpus.

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*, Brussels, Belgium.

Jue Wang, Lidan Shou, Ke Chen, and Gang Chen. 2020. Pyramid: A Layered Model for Nested Named Entity Recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5918–5928, Online.

Chih-Hsuan Wei, Yifan Peng, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Jiao Li, Thomas C Wiegers, and Zhiyong Lu. 2016. Assessing the state of the art in biomedical relation extraction: overview of the BioCreative V chemical-disease relation (CDR) task. *Database*, 2016.

Shanchan Wu and Yifan He. 2019. Enriching pretrained language model with entity information for relation classification. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 2361–2364.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. LUKE: Deep Contextualized Entity Representations with Entity-aware Self-attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online.

Juntao Yu, Bernd Bohnet, and Massimo Poesio. 2020. Named Entity Recognition as Dependency Parsing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6470–6476, Online.

Zexuan Zhong and Danqi Chen. 2021. A Frustratingly Easy Approach for Entity and Relation Extraction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 50–61, Online.