Biomedical NER using Novel Schema and Distant Supervision

Anshita Khandelwal, Alok Kar, Veera Raghavendra Chikka and **Kamalakar Karlapalem**
Data Science and Analytics Center (DSAC),
Kohli Center for Information Systems (KCIS),
International Institute of Information Technology, Hyderabad (IIIT-H)
{anshita.khandelwal, alok.kar, raghavendra.ch}@alumni.iiit.ac.in,
kamal@iiit.ac.in

Abstract

Biomedical Named Entity Recognition (BMNER) is one of the most important tasks in the field of biomedical text mining. Most work so far on this task has not focused on identification of discontinuous and overlapping entities, even though they are present in significant fractions in real-life biomedical datasets. In this paper, we introduce a novel annotation schema to capture complex entities, and explore the effects of distant supervision on our deep-learning sequence labelling model. For BMNER task, our annotation schema outperforms other BIO-based annotation schemes on the same model. We also achieve higher F1-scores than state-of-the-art models on multiple corpora without fine-tuning embeddings, highlighting the efficacy of neural feature extraction using our model.

1 Introduction

Named entity recognition (NER) consists of identification and classification of named entities in text. Biomedical NER (BMNER) is a crucial problem in healthcare as it is the initial step in solving various tasks, such as relation extraction, semantic role labeling, and clinical decision making (De Bruijn and Martin, 2002)(Hanisch et al., 2003). As compared to NER in other domains, BMNER is a difficult task as labelled data in biomedical domain is less in amount and expensive to obtain, and it requires identification of complex entities that are not common in other domains (Dai, 2018). Recently, deep learning approaches using large unstructured data, such as Bi-LSTM with CRF (Li et al., 2018) and BERT (Symeonidou et al., 2019)(Yu et al., 2019) models have been used to obtain state-of-the-art results on BMNER.

The most common annotation scheme for NER is BIO tagging, where B is for Beginning of entity, I for Inside of entity, and O for Outside of entity. A major assumption of BIO tagging is that an entity is composed of continuous and non-overlapping tokens. As complex entities that defy these assumptions frequently occur in biomedical records, a new scheme is needed to capture them. For this purpose, BIOHD (Tang et al., 2013) was introduced to represent discontinuous entities that may overlap with four new tags : (BH,IH) as shared head tags and (BD,ID) as non-shared non-head tags. However, this scheme fails to capture discontinuous entities that have more than two spans. In this paper, we propose a novel annotation schema BIODT that overcomes this limitation of BIOHD. Our schema includes shared non-head tags and non-shared head tags, and hence captures entities with more than two spans, which BIOHD fails to do.

Distant supervision is a method to generate labelled data from unlabelled data using existing knowledge (Mintz et al., 2009) that is particularly useful to create data for supervised learning algorithms which require large amounts of data. We use this method for BMNER to compensate for the lack of labelled data in the biomedical domain. As we are using an RNN (Recurrent Neural Network) which requires a large amount of training data, distant supervision helps in increasing the amount of annotated records without human effort.

In summary, the main contributions of this paper are as follows:

1. A novel systematic tagging schema to better capture discontinuous entities, that is significantly better($>2\%$) for prediction of discontinuous entities than BIOHD.
2. A distant supervision approach to biomedical NER, that uses labelled data to generate labels for unlabelled data, without the use of external dictionaries. Our experiments show that distant supervision methods boost the performance of our model, and also outperform state-of-the-art models.

Proceedings of the BioNLP 2022 workshop, Dublin, Ireland, pages 155–160

2 Related Work

Existing solutions for BMNER include traditional NER methods such as dictionary or rule-based approaches, as well as supervised machine learning methods like Markov models (Ponomareva et al., 2007), Conditional Random Fields (CRFs) (Ponomareva et al., 2007)(Sun et al., 2006)(Settles, 2004) and Support Vector Machine (SVM) (Ju et al., 2011)(Kazama et al., 2002). Lately, deep learning approaches using large unstructured data, such as Bi-LSTM with CRF (Li et al., 2018) and BERT (Symeonidou et al., 2019)(Yu et al., 2019) models have been used to obtain state-of-the-art results on BMNER. To deal with scarcity of token-level annotated data required in deep-learning models, some weak-supervision and distant-supervision solutions have been proposed. For the task of BMNER, Mathew et al. (Mathew et al., 2019) introduced a weakly-supervised data augmentation approach for identification of proteins in BioCreative Challenge VI Track 1 dataset(Arighi et al., 2018), using a reference set of entity names from knowledge bases like UniProt (Consortium, 2018) to identify entity mentions on unlabelled data. In 2016, Lee et al. proposed a bagging-based approach using active learning with distant supervision, that uses a semi-automatically constructed dictionary of named entities from Wikipedia (Lee et al., 2016) (Song and Kim, 2015). To the best of our knowledge, no prior work has been done to study the effects of distant supervision on complex entities for NER.

To deal with annotation of complex entities, many methods have been proposed. Annotation schemes like BIOHD (Tang et al., 2013) and BIOHD1234 (Tang et al., 2015) were proposed with four and ten additional tags, respectively, to the commonly used BIO schema. These schemes gave near state-of-the-art results with simple machine learning models. Methods such as representing sentences as hypergraphs (Lu and Roth, 2015) (Muis and Lu, 2016), transition-based models that uses specialized actions and attention mechanisms (Dai et al., 2020), and representing NER task as a structured multi-label classification problem (McDonald et al., 2005) have also been explored. Additionally, a two-stage approach that first detects all continuous parts, then combines them to form discontiguous entities using a classifier (Wang and Lu, 2019) has also been proposed.

3 Annotation Schema

We introduce a new annotation schema called BIODT, which consists of 11 tags: the traditional BIO tags, and 8 additional tags as described below.

1. DB, DI are shared heads of the first term in a discontinuous entity
2. DHB, DHI are shared non-head tags of the subsequent terms in a discontinuous entity
3. TB, TI are non-shared heads of the first term in a discontinuous entity
4. THB, THI are non-shared non-head tags of the subsequent terms in a discontinuous entity

Preference is given to combine shared head tags with shared non-head tags and, similarly, for non-shared tags. For example, in sentence 1 (Figure 1), "aortic root", "descending root" and "dilated" are tagged with shared tags. Similarly, in Sentence 2, "mitral", "leaflet" and "thickened" are tagged with non-shared tags. There are a few cases where shared and non-shared tags can co-occur in a sentence. In Sentence 3 (Figure 1), "ABD" is a shared head tag. If tagged according to BIOHD schema, "tenderness" and "RUQ" would be shared non-head tags, resulting in two entities, "ABD...tenderness" and "ABD...RUQ", which are wrong. In our schema, we tag "tenderness" and "RUQ" with non-shared non-head tags(TH{B,I}), which are combined with the shared head tag(D{B,I}) to form "ABD...tenderness...RUQ". Hence, our schema captures entities that were not captured by the BIOHD schema.

Extracting entities from BIODT tagged sentence:
Discontinuous entities can be obtained from a BIODT tagged sentence using the following simple rules :

- For shared tags :
  1. Each shared non-head tag (DH{B,I}) is joined to each shared head tag (D{B,I}) in the sentence.
  2. If no shared head tag is present, all shared non-head tags in the sentence are combined to form one joined entity.
- For non-shared tags :
  1. All non-shared non-head tags (TH{B,I}) are joined together.
  2. If any non-shared head tag (T{B,I}) is present, then the entity obtained from (1) is joined to each non-shared head tag in the sentence.

3. If no non-shared head tag is present and any shared head tag (D{B,I}) is present, then the entity obtained from (1) is joined to each shared head tag in the sentence.
4. If no head tags(shared/non-shared) are present in the sentence, return entity obtained from (1).

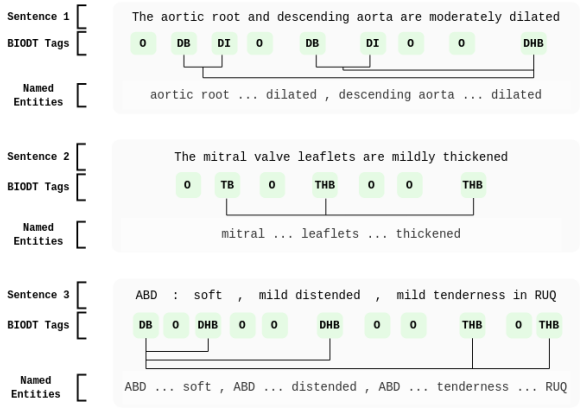

Figure 1: BIODT Schema Examples

4 Approach and Architecture

We use a BiLSTM-CRF network to assign labels for NER, as presented in Figure 2. BiLSTM-CRF is an RNN (Recurrent Neural Network), and is formed by the combination of a BiLSTM (Bidirectional Long-Short Memory) and a CRF (Conditional Network Field). For each sentence, the BiLSTM forms a vector representation for each word, preserving backward and forward context. This vector representation is then used as the input to the CRF, which predicts labels for the words of the sentence.

The labels at the CRF output layer are decoded using the Viterbi algorithm.

4.1 Features and Embeddings

We have used a combination of GloVe word embeddings(Pennington et al., 2014), character embeddings and BERT (Bio+Discharge Summary BERT) embeddings (Alsentzer et al., 2019). Additionally, we have also experimented with part-of-speech(POS) embeddings, case(lower/upper) embeddings, and suffix/prefix embeddings.

4.2 Distant Supervision

We use unlabelled data to generate a larger training set using distant supervision. We trained our baseline model on manually annotated data, then used the model to predict labels on additional unlabelled data to expand our training set.

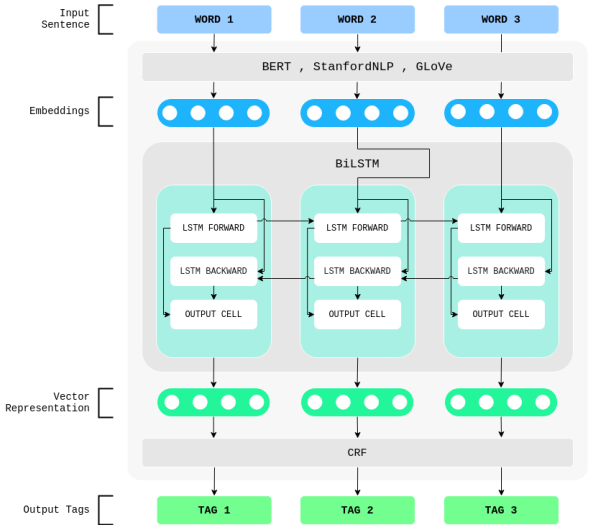

Figure 2: Model Architecture

Our method is 1. train M on D1-train, 2. predict labels on unlabelled dataset D2 and augment this newly labelled dataset to D1-train, 3. train M on D1-train and newly labelled D2, 4. finally, we test M on D1-test.

Here, D1-train and D1-test are train and test partitions of the labelled dataset respectively, D2 is an unlabelled dataset of similar domain, and M is our model.

5 Datasets

We experiment on two datasets from the biomedical domain: ShARe 2013 (Forner et al., 2013) and ShARe 2014 (Cappellato et al., 2014). The datasets contain clinical free-text notes, which include discharge summaries, echo-reports, and radiology reports. An annotated named entity can contain any number of continuous spans, and it maps to a concept in the disorder semantic group of SNOMED-CT (Cornet and de Keizer, 2008).

Since a significant fraction(almost 10%) of the mentions in these datasets are discontinuous(from Table 1), an improvement in discontinuous entity recognition will show noticeable improvement in overall entity recognition.

	ShARe 2013	ShARe 2014
#Records	298	433
#Sentences	18.7k	34.6k
#Total Mentions	11,161	19,131
#Disc. Mentions	1,090	1,710
% Disc. Mentions	9.7	8.9

Table 1: Dataset statistics for ShARe 2013 and ShARe 2014

6 Results and Analysis

For evaluation, we have used scripts provided in ShARe tasks to calculate F-score (F) to evaluate the efficiency of the models in our experimentation.

Our baseline model is a BiLSTM-CRF that uses the features and embeddings mentioned in 4.1, as proposed by Yu et al. (Yu et al., 2019).

We faced a replication crisis while attempting to reproduce the results presented in (Tang et al., 2015) using the proposed BIOHD1234 schema. Hence, we were unable to compare the performance of our schema with that of BIOHD1234.

6.1 Model Evaluation

As can be seen from Table 2, our model outperforms the baseline in both annotation schemes by a small margin. It also gives a better result than the state-of-the-art by 1.6% and 1.1% for both datasets, using BIODT and BIOHD schemes, respectively. Evaluating for discontinuous entities, we find that our model performance is similar to that of the baseline task, with the BIOHD schema slightly underperforming for the ShARe 2013 corpus.

Model	Scheme	ShARe 13	ShARe 14
SSVM (Tang et al., 2013)	BIOHD	75.0	–
SSVM (Tang et al., 2015)	BIOHD1234	78.3	–
Transition-based model (Dai et al., 2020)	HGB	77.7	79.6
Baseline	BIOHD	78.4	79.7
Distant Supervision	BIOHD	78.9	**80.7**
Baseline	BIODT	79.0	80.4
Distant Supervision	BIODT	**79.9**	80.5

Table 2: F1-Scores of other models compared to our model; HGB stands for Hypergraph Based

Dataset	Model	BIOHD	BIODT
ShARe 2013	(Tang et al., 2015)	48.7	-
	Baseline	46.1	51.6
	D. Supervision	45.6	**52.8**
ShARe 2014	Baseline	40.5	44.2
	D. Supervision	41.9	**44.5**

Table 3: F1-Scores with BIOHD and BIODT for discontinuous entities

6.2 Evaluation of Annotation Schema

On entire datasets, BIODT performs similar(within 1%) to BIOHD for all models. The only case where it is not an improvement over BIOHD is when we use our model on ShARe 2014 dataset, where it has a 0.2% less score. As is clear from Table 3, BIODT schema gives a significantly better performance over BIOHD for discontinuous entities(>3%), for all cases.

6.3 Analysis

From Table 2 and Table 3, it can be inferred that while BIODT does not help much for NER in entire datasets, it brings a noticeable improvement compared to BIOHD for discontinuous entities. We believe that for datasets with a higher fraction of discontinuous entities, BIODT will perform better than it has for these experiments.

From Table 2 and Table 3, it is also clear that when used with BIODT schema, distant supervision enhances performance, both for entire datasets and for discontinuous entities.

Limitations of BIODT

Due to the decoding rules of BIODT, some false positives occur even on correctly predicted labels:

```
DB1 DI1 O O DB2 DI2 O O O
DHB1 DHI1 O O O DHB2 DHI2
```

Here, the original entities are :

```
(DB1 DI1 DHB1 DHI1) , (DB2 DI2 DHB2 DHI2)
```

Now, according to decoding rules, each shared non-head term will combine to each shared head term, hence the entities obtained will be :

1. `DB1 DI1 DHB1 DHI1`
2. `DB1 DI1 DHB2 DHI2`
3. `DB2 DI2 DHB1 DHI1`
4. `DB2 DI2 DHB2 DHI2`

Among these entities, (1) and (4) are correctly decoded, (2) and (3) are not. Even if our model predicts these labels correctly, they will be decoded as false positives. We do not believe that this leads to worse performance of BIODT as compared to BIOHD, as BIOHD faces a similar problem.

7 Conclusion

In this paper, we introduced a novel annotation schema to identify named entities in biomedical data. We have also shown that for the same model, our annotation scheme gives better performance than other BIO-based complex annotation schemes for discontinuous entities. We also explore the distant supervision paradigm to increase our training set for BioNER. Using this, we have achieved state-of-the-art results.

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Cecilia Arighi, Lynette Hirschman, Thomas Lemberger, Samuel Bayer, Robin Liechti, Donald Comeau, and Cathy Wu. 2018. Bio-id track overview. In *Proceedings of the BioCreative VI Workshop*, pages 14–19, Bethesda, MD, USA.

Linda Cappellato, Nicola Ferro, Martin Halvey, and Wessel Kraaij, editors. 2014. *Working Notes for CLEF 2014 Conference, Sheffield, UK, September 15-18, 2014*, volume 1180 of *CEUR Workshop Proceedings*. CEUR-WS.org.

The UniProt Consortium. 2018. Uniprot: a worldwide hub of protein knowledge. *Nucleic Acids Research*, 47(D1):D506–D515.

Ronald Cornet and Nicolette de Keizer. 2008. Forty years of snomed: a literature review. *BMC medical informatics and decision making*, 8 Suppl 1:S2.

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

Xiang Dai, Sarvnaz Karimi, Ben Hachey, and Cecile Paris. 2020. An effective transition-based model for discontinuous NER. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5860–5870, Online. Association for Computational Linguistics.

Berry De Bruijn and Joel Martin. 2002. Getting to the (c) ore of knowledge: mining biomedical literature. *International journal of medical informatics*, 67(1):7–18.

Pamela Forner, Roberto Navigli, Dan Tufis, and Nicola Ferro, editors. 2013. *Working Notes for CLEF 2013 Conference , Valencia, Spain, September 23-26, 2013*, volume 1179 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Daniel Hanisch, Juliane Fluck, Heinz-Theodor Mevissen, and Ralf Zimmer. 2003. Playing biology's name game: Identifying protein names in scientific text. In *Proceedings of the 8th Pacific Symposium on Biocomputing, PSB 2003, Lihue, Hawaii, USA, January 3-7, 2003*, pages 403–414.

Z. Ju, J. Wang, and F. Zhu. 2011. Named entity recognition from biomedical text using svm. In *2011 5th International Conference on Bioinformatics and Biomedical Engineering*, pages 1–4.

Jun'ichi Kazama, Takaki Makino, Yoshihiro Ohta, and Jun'ichi Tsujii. 2002. Tuning support vector machines for biomedical named entity recognition. In *Proceedings of the ACL-02 Workshop on Natural Language Processing in the Biomedical Domain*, pages 1–8, Phildadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Sunghee Lee, Yeongkil Song, Maengsik Choi, and Harksoo Kim. 2016. Bagging-based active learning model for named entity recognition with distant supervision. In *Proceedings of the 2016 International Conference on Big Data and Smart Computing (BigComp)*, BIGCOMP '16, page 321–324, USA. IEEE Computer Society.

Fei Li, Meishan Zhang, Bo Tian, Bo Chen, Guohong Fu, and Donghong Ji. 2018. Recognizing irregular entities in biomedical text via deep neural networks. *Pattern Recognition Letters*, 105(C):105–113.

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. Association for Computational Linguistics.

Joel Mathew, Shobeir Fakhraei, and José Luis Ambite. 2019. Biomedical named entity recognition via reference-set augmented bootstrapping.

Ryan McDonald, Koby Crammer, and Fernando Pereira. 2005. Flexible text segmentation with structured multilabel classification. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 987–994, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2*, ACL '09, page 1003–1011, USA. Association for Computational Linguistics.

Aldrian Obaja Muis and Wei Lu. 2016. Learning to recognize discontiguous entities. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 75–84, Austin, Texas. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. *Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014)*, 12:1532–1543.

Natalia Ponomareva, Paolo Rosso, Ferran Pla, and Antonio Molina. 2007. Conditional random fields vs. hidden markov models in a biomedical named entity recognition task.

Burr Settles. 2004. Biomedical named entity recognition using conditional random fields and rich feature sets. JNLPBA '04, page 104–107, USA. Association for Computational Linguistics.

Yeongkil Song and Harksoo Kim. 2015. Semi-automatic construction of a named entity dictionary based on active learning. In *Computer Science and its Applications*, pages 65–70, Berlin, Heidelberg. Springer Berlin Heidelberg.

Chengjie Sun, Yi Guan, Xiaolong Wang, and Lei Lin. 2006. Biomedical named entities recognition using conditional random fields model. In *Fuzzy Systems and Knowledge Discovery*, pages 1279–1288, Berlin, Heidelberg. Springer Berlin Heidelberg.

Anthi Symeonidou, Viachaslau Sazonau, and Paul Groth. 2019. Transfer learning for biomedical named entity recognition with biobert. In *Posters and Demo Track of the 15th International Conference on Semantic Systems. (Poster and Demo Track at SEMANTiCS 2019)*, number 2451 in CEUR Workshop Proceedings, pages 126–130, Aachen.

Buzhou Tang, Qingcai Chen, Xiaolong Wang, Yonghui Wu, Yaoyun Zhang, Min Jiang, Jingqi Wang, and Hua Xu. 2015. Recognizing disjoint clinical concepts in clinical text using machine learning-based methods. *AMIA Annual Symposium Proceedings*, 2015:1184–1193.

Buzhou Tang, Yonghui Wu, Min Jiang, Joshua C Denny, and Hua Xu. 2013. Recognizing and encoding discorder concepts in clinical text using machine learning and vector space model. In *Workshop of ShARe/CLEF eHealth Evaluation Lab*.

Bailin Wang and Wei Lu. 2019. Combining spans into entities: A neural two-stage approach for recognizing discontiguous entities. 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 6216–6224, Hong Kong, China. Association for Computational Linguistics.

X. Yu, W. Hu, S. Lu, X. Sun, and Z. Yuan. 2019. Biobert based named entity recognition in electronic medical record. In *2019 10th International Conference on Information Technology in Medicine and Education (ITME)*, pages 49–52, Los Alamitos, CA, USA. IEEE Computer Society.