FuzzyBIO: A Proposal for Fuzzy Representation of Discontinuous Entities

Anne Dirkson
LIACS, Leiden University
Niels Bohrweg 1, 2333 CA,
Leiden, the Netherlands

Suzan Verberne
LIACS, Leiden University
Niels Bohrweg 1, 2333 CA
Leiden, the Netherlands

Wessel Kraaij
LIACS, Leiden University
Niels Bohrweg 1, 2333 CA
Leiden, the Netherlands

{a.r.dirkson, s.verberne, w.kraaij}@liacs.leidenuniv.nl

Abstract

Discontinuous entities pose a challenge to named entity recognition (NER). These phenomena occur commonly in the biomedical domain. As a solution, expansions of the BIO representation scheme that can handle these entity types are commonly used (i.e. BIOHD). However, the extra tag types make the NER task more difficult to learn. In this paper we propose an alternative; a fuzzy continuous BIO scheme (FuzzyBIO). We focus on the task of Adverse Drug Response extraction and normalization to compare FuzzyBIO to BIOHD. We find that FuzzyBIO improves recall of NER for two of three data sets and results in a higher percentage of correctly identified disjoint and composite entities for all data sets. Using FuzzyBIO also improves end-to-end performance for continuous and composite entities in two of three data sets. Since FuzzyBIO improves performance for some data sets and the conversion from BIOHD to FuzzyBIO is straightforward, we recommend investigating which is more effective for any data set containing discontinuous entities.

1 Introduction

Adverse Drug Reactions (ADRs), harmful reactions that result from the intake of medication, pose a major health concern (World Health Organisation, 2006). Due to the limitations of clinical trials on the one hand (Shenoy and Harugeri, 2015) and reporting systems after release on the market on the other hand (Hazell and Shakir, 2006), many ADRs remain undiscovered. Therefore, both social media and clinical reports are being explored by the research community as alternative information sources for the semi-automatic discovery of ADRs (Lardon et al., 2015; Sarker et al., 2015).

One particular challenge for the extraction of ADRs from text is the presence of discontinuous entities. These can be either composite entities (i.e. some words belong to multiple entities), such as '*lack of sleep* and *appetite*', or disjoint entities (i.e. split entities), such as '*eyes* are feeling *dry*'. These phenomena occur more commonly in the clinical than general domain. In fact, Tang et al. (2015) reported that discontinuous mentions in clinical text account for about 10% of all ADR mentions. None of the traditional versions of the BIO representation scheme (B: beginning of entity, I: inside entity and O: outside entity) or common extensions such as IOBES (E: end of entity, S: singleton entity)[1] were designed to handle such mentions (Pradhan et al., 2014). Therefore, Tang et al. (2015) proposed extending the BIO scheme with two additional tags: the 'H' for words shared by multiple mentions and 'D' for parts of discontinuous mentions not shared by other mentions. This resulted in four new tag types (HB-, HI-, DB- and DI-). Their BIOHD representation was broadly adopted by the community (Karimi et al., 2015; Zolnoori et al., 2019; Si et al., 2019). Table 1 shows examples of concepts represented with the BIOHD scheme.

Although the BIOHD scheme allows for precise representation of entities, the extra tag types make the task more difficult for models to learn. Straightforward BIO rules such as 'an entity always starts with a B' are no longer valid under the BIOHD scheme. In this paper we argue that a more simple BIO representation in which discontinuous entities are transformed into continuous sequences by including all non-entity tokens in between would improve ADR extraction by being easier to learn and reintroducing these straightforward rules. We coin this representation FuzzyBIO. Some examples of entities represented with BIOHD and FuzzyBIO can be seen in Table 1.

Aside from improving extraction, using FuzzyBIO instead of BIOHD may also improve sub-

[1] This scheme is also called BIOES, BILOU or BMEWO

Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis, pages 77–82
April 19, 2021.

Sentence 1	**Muscles**	are	constantly	**quivering**	!				
BIOHD	DB	O	O	DI	O				
FuzzyBIO	B	I	I	I	O				
Sentence 2	I	have	**pain**	**in**	**my**	**hands**	and	**upper**	**arms**
BIOHD	O	O	HB	HI	HI	DB	O	DB	DI
FuzzyBIO	O	O	B	I	I	I	I	I	I

Table 1: Examples of discontinuous disjoint (sentence 1) and composite (sentence 2) ADR mentions represented by the BIOHD and FuzzyBIO schemes.

sequent concept normalization, in which ADRs are linked to standardized medical concepts (e.g. 'can't fall asleep' to the concept 'insomnia' with concept identifier 193462001 in the medical ontology SNOMED-CT). This step is essential for aggregating and thus quantifying the prevalence of ADR. As current normalization methods are mostly hampered by errors made during extraction (Weissenbacher et al., 2019), this step may also benefit from simplification of the representation scheme.

Thus, we address two research questions:

RQ1 To what extent can fuzzy continuous representation of discontinuous entities improve NER of ADR? (intrinsic evaluation)

RQ2 To what extent can this fuzzy representation benefit end-to-end ADR extraction? (extrinsic evaluation)

In this paper we present FuzzyBIO, a fuzzy continuous BIO representation of discontinuous entities. Moreover, we show it is beneficial for end-to-end ADR discovery. Our representation is applicable to other domains as well. We release our code for the purpose of follow-up research.[2]

2 Related Work

The first shared task to deal with discontinuous medical entities was SemEval 2014 Task 7 (Pradhan et al., 2014). Prior to this task, discontinuous entities were often excluded (e.g. Uzuner et al. (2011)) or each part was represented as a separate continuous entity and later reassembled (Metke-Jimenez and Karimi, 2015). Various representations were proposed but the only one able to distinguish between those that share a head word (i.e composite entities) and those that do not (e.g. disjoint entities) was the BIOHD scheme (Zhang et al., 2014).

This scheme was later analysed in more detail and compared to two baseline approaches: (1) ignoring all discontinuous entities and (2) representing separate parts of discontinuous entities as individual entities (Tang et al., 2015). In comparison to the baseline approaches, the BIOHD scheme could improve recognition of both discontinuous but also continuous entities, likely due to its ability to distinguish between the two.

Tang et al. (2015) also proposed a further extension (BIOHD1234) in which numbers were added to refer to which entity a non-head ('D') entity should be combined with[3], effectively expanding the scheme from 7 to 13 tags. This representation was able to outperform BIOHD due to its ability to correctly represent multiple discontinuous entities and discontinuous entities with more than one non-head part. However, as neither BIOHD nor BIOHD1234 could handle multiple head entities in one sentence, Tang et al. (2018) proposed a multi-label BIO representation in which tokens can be labeled with more than one tag, and each tag corresponds to one entity. For NER of adverse drug responses, this novel representation managed to outperform BIOHD. Similarly, Shang et al. (2020) allowed for multiple labels per token for extracting disorders from scientific articles.

Despite its limitations (Tang et al., 2018), the BIOHD scheme is commonly adopted (Si et al., 2019; Karimi et al., 2015; Zolnoori et al., 2019). We propose an alternative, simpler representation scheme that could improve extraction by being easier to learn.

3 Methods

3.1 The FuzzyBIO representation scheme

As displayed in Table 1, FuzzyBIO transforms discontinuous into continuous entities by annotating

[2]Code is available at: https://github.com/AnneDirkson/FuzzyBIO

[3]1 and 2 denote nearest head and non-head entity on the left, and 3 and 4 denote nearest head and non-head entity on the right

all tokens in between.[4] Composite entities are combined if they share an entity head. We realise this compresses two separate entities into one (e.g. the entities 'pain in my hands' and 'pain in my upper arms' in Table 1). However, this does not pose a problem to normalization, as the state-of-the-art normalization method (Sung et al., 2020) includes heuristic rules to split composite entities prior to normalization.

3.2 Named Entity Recognition of ADR

For the NER task itself, we opt for distilBERT (base-cased), a lighter more computationally efficient version of BERT (Sanh et al., 2019). We use a one-cycle learning rate (LR) policy (Smith, 2018) with a maximum LR of 0.01. For each fold in the 10-fold cross-validation (CV), we select either 3 or 4 epochs based on the validation data. We use the Huggingface implementation (Wolf et al., 2019) with the wrapper ktrain (Maiya, 2020) to train our models with the initialization seed set to 1.

3.3 Concept normalization of ADR

For normalization, we use the state-of-the-art BioSyn method with default parameters (Sung et al., 2020; Tutubalina et al., 2020). It is possible to provide composite entities as input, as this method splits composite entities prior to normalization using the heuristics by Souza and Ng (2015).

Our target ontology is SNOMED-CT[5]. As SNOMED-CT is too extensive for our purpose, we aim to map SNOMED concepts in our training data to a curated subset of SNOMED, the CORE Problem List Subset[6], before training the normalization model. If there is a direct mapping in the community based mappings in BioPortal (Noy et al., 2008) between the original concept and a CORE concept or the parent of the concept is in the CORE (e.g. 'moderate anxiety' to 'anxiety'), we map the mention to the respective CORE concept. We include all concepts of the CORE subset and all concepts that could not be mapped to a CORE concept in the data as candidates. Synonyms for each concept are retrieved from the community based mappings in BioPortal (Noy et al., 2008) using the REST API and from the UMLS using pymedtermino (Lamy et al., 2015).

[4] In our data, we did not find any cases where an entity was lost because it was in between two parts of another discontinuous entity. Nonetheless, this is theoretically possible and poses a potential limitation.

[5] https://www.nlm.nih.gov/healthit/snomedct/index.html

[6] https://www.nlm.nih.gov/research/umls/Snomed/core_subset.html

Entities	CADEC	PsyTAR	CLEF
Continuous	5.360	4.508	16.261
Discontinuous			
– Disjoint	100	225	909
– Composite	828	70	286

Table 2: Size of data sets.

3.4 Evaluation

For evaluating the NER models on a token level, our metrics are lenient and ignore the prefixes (B- I- H- D-). Additionally, we evaluate performance on an entity level. Following Magge et al. (2020), an entity is considered a true positive if any part of the annotated ADR text is correctly identified (i.e. overlaps with the predicted ADR text). We evaluate the end-to-end performance by calculating how many entities were both extracted during NER and normalized to the correct SNOMED-CT concept.

4 Data

We use two data sets of social media posts annotated for ADR: CADEC (Karimi et al., 2015) and PsyTAR (Zolnoori et al., 2019). The former was also used by Tang et al. (2018). Both contain posts from medical fora on AskaPatient.com. Additionally, we used a data set of clinical records annotated for disorder mentions, namely the SemEval 2014 Task 7 data (Pradhan et al., 2014) that builds on the CLEF eHealth 2013 corpus used by Tang et al. (2015). See Table 2 for more details. Data sets were split into 10 folds stratified on the presence of ADRs. For PsyTAR, we chose sentences as units as they were annotated separately. For the CLEF data set, each document was split into sequences of 5 sentences for the NER task, because of memory restrictions on the input length for BERT models.

5 Results

5.1 Intrinsic evaluation

As can be seen in Table 3, the FuzzyBIO scheme improves recall for two of three data sets, namely for CADEC (+0.29) and CLEF (+0.07), at a cost to precision (-0.24 and -0.1). For these data sets, using FuzzyBIO also leads to a higher percentage of correctly identified entities for both continuous (+1.1 and +0.6) and discontinuous entities (+3.5

Data	Scheme	Token level			Entity level recall		
		Micro F_1	P	R	Continuous	Disjoint	Composite
CADEC	BIOHD	0.586	**0.636**	0.555	42.0%	55.4%	64.6%
	FuzzyBIO	**0.596**	0.612	**0.584**	**43.1%**	**58.9%**	**65.2%**
PsyTAR	BIOHD	**0.771**	**0.751**	**0.797**	**87.4%**	83.8%	79.5%
	FuzzyBIO	0.762	0.747	0.780	84.9%	**84.4%**	**87.5%**
CLEF	BIOHD	**0.312**	**0.286**	0.345	35.6%	52.8%	69.1%
	FuzzyBIO	0.309	0.276	**0.352**	**36.2%**	**55.2%**	**76.6%**

Table 3: Intrinsic evaluation of NER. Results are the average of a 10 fold CV.

and +3.6 for disjoint and +0.6 and +0.7 for composite entities). For the remaining data set (PsyTAR), overall NER performance is negatively affected by using FuzzyBIO (-0.09) and continuous entities are missed more often (-2.5). Nonetheless, also for this data set discontinuous entities are extracted correctly more often (+0.8 and +8.0) when using FuzzyBIO instead of the BIOHD scheme.

5.2 Extrinsic evaluation

As can be seen in Table 4, using the FuzzyBIO scheme improves end-to-end performance for continuous and composite entities in two of three data sets, namely the CADEC (+0.4 and +1.3) and CLEF data (+0.4 and +5.7). In contrast, the end-to-end performance for disjoint entities is decreased (-15.5 and -21.0) for these data sets despite initial gains during NER. In the remaining data set (PsyTAR), the percentage of correctly identified entities after normalization is lower for all entity types when using the FuzzyBIO instead of the BIOHD scheme.

Data	Scheme	Entity level recall		
		Continuous	Disjoint	Composite
CADEC	BIOHD	23.9%	**35.9%**	21.2%
	FuzzyBIO	**24.3%**	20.4%	**22.5%**
PsyTAR	BIOHD	**43.6%**	**26.1%**	**10.6%**
	FuzzyBIO	42.8%	25.0%	7.5%
CLEF	BIOHD	21.7%	**25.8%**	26.5%
	FuzzyBIO	**22.1%**	4.8%	**32.2%**

Table 4: Extrinsic evaluation of ADR extraction. Results are the average of a 10 fold CV.

6 Discussion

In answer to RQ1, we find that the FuzzyBIO scheme benefits overall recall during NER for two of the three data sets. For these data sets, it also leads to a higher percentage of correctly identified entities, both continuous and discontinuous. For the third data set (PsyTAR), more discontinuous entities are extracted correctly when using FuzzyBIO compared to the BIOHD scheme. However, more continuous entities are missed. In answer to RQ2, we find that for the same two data sets (CADEC and CLEF) the end-to-end ADR extraction is improved for continuous and composite entities. However, for the remaining data set (PsyTAR) the end-to-end performance is lower for all entity types.

We believe that the difference between PsyTAR and the other data sets may be related to either the low number of discontinuous entities or the low number of composite entities in the PsyTAR data, which may have hindered the training of an NER model for these entity types. An alternative explanation is that FuzzyBIO is less beneficial for easier NER tasks: The initial NER performance with BIOHD is far higher for PsyTAR (F_1 of 0.771) than for the other data sets (F_1 of 0.586 and 0.312). The difference between PsyTAR and the other data sets is unlikely to be related to the relative percentage of discontinuous entities, as this is similar to that of the CLEF data (5.8 vs 6.2%), or the nature of the data, as CADEC contains forum posts from the same website.

Another result that stands out is the lower end-to-end performance for disjoint entities when using the FuzzyBIO scheme despite initial gains in the extraction of disjoint entities for all data sets. We suspect that normalization of these entities is made more challenging by the words in between the disjoint parts of the entity that are now included in the extracted entity. Therefore, in future work, we plan to investigate post-processing steps such as the removal of stop words which may improve the normalization of the more noisy disjoint entities represented with FuzzyBIO. As our representation is applicable for representing any type of discontinuous entity, future work may also include testing FuzzyBIO in other domains.

Although the improvement in NER is compa-

rable for disjoint entities in medical records and user-generated content, the negative impact on normalization of disjoint entities is far stronger for the medical records. One might expect the normalization to decrease more strongly if more non-entity words (i.e. more noise) were included, but the median amount of non-entity words included in the disjoint entities is equal for all data sets (on average 1 word is added). Manual analysis also does not reveal a difference between the *type* of non-entity word included; They appear to mostly be stopwords. Thus, the most likely explanation for this difference is that the training examples for normalization from the user-generated data are already more noisy than their counterparts from the medical records. Consequently, the normalization algorithm for user-generated data might be better at dealing with noise. Future work could investigate whether training with the noisy examples instead of the original entities would be beneficial.

FuzzyBIO appears to be more beneficial for end-to-end extraction of composite entities in the medical records (+5.7) than in the user-generated data (+1.3 and -3.1). However, the number of non-entity words that is included is not lower for the medical records (median of 2 words added) compared to the user-generated data (median of 1 for CADEC and 3 for PsyTAR). Thus, this difference does not appear to be due to an increase in the fuzziness of the entities.

We also find some support for our hypothesis that the BIOHD representation makes the NER task more difficult for BERT models to learn than the FuzzyBIO representation. Overall the BERT models have difficulty learning the additional tag types; The precision for H- and D-tags is consistently lower than the precision for B-tags. In fact, on the PsyTAR data which contains only few overlapping entities (1.7%), the H-tag was never predicted. It seems that FuzzyBIO makes the task easier in two ways, namely by standardizing entities into continuous sequences that always start with a B-tag and by excluding rare tags such as the H-tag. Standardizing entities makes it easier for the model to learn the underlying rules and excluding rare tags removes a goal for which there are only few examples available.

7 Conclusion

As recommendation for future NER tasks, we expect FuzzyBIO to especially benefit NER for difficult tasks with a fair amount of discontinuous entities. However, since the conversion from BIOHD to FuzzyBIO is straightforward and deterministic, we recommend to experimentally compare which of the two representations is more effective for any data set that includes discontinuous entities.

Acknowledgments

We would like to thank the SIDN fonds for funding this research and our reviewers for their valuable feedback.

References

Lorna Hazell and Saad A W Shakir. 2006. Under-Reporting of Adverse Drug Reactions A Systematic Review. *Drug Safety*, 29(5):385–396.

Sarvnaz Karimi, Alejandro Metke-Jimenez, Madonna Kemp, and Chen Wang. 2015. Cadec: A corpus of adverse drug event annotations. *Journal of Biomedical Informatics*, 55:73–81.

Jean Baptiste Lamy, Alain Venot, and Catherine Duclos. 2015. PyMedTermino: An open-source generic API for advanced terminology services. *Studies in Health Technology and Informatics*, 210:924–928.

Jérémy Lardon, Redhouane Abdellaoui, Florelle Bellet, Hadyl Asfari, Julien Souvignet, Nathalie Texier, Marie Christine Jaulent, Marie Noëlle Beyens, Anita Burgun, and Cédric Bousquet. 2015. Adverse drug reaction identification and extraction in social media: A scoping review. *Journal of Medical Internet Research*, 17(7):1–16.

Arjun Magge, Elena Tutubalina, Zulfat Miftahutdinov, Ilseyar Alimova, Anne Dirkson, Suzan Verberne, Davy Weissenbacher, and Graciela Gonzalez-Hernandez. 2020. Deepademiner: A deep learning pharmacovigilance pipeline for extraction and normalization of adverse drug effect mentions on twitter. *medRxiv*.

Arun S. Maiya. 2020. ktrain: A low-code library for augmented machine learning. *arXiv*, arXiv:2004.10703 [cs.LG].

Alejandro Metke-Jimenez and Sarvnaz Karimi. 2015. Concept Extraction to Identify Adverse Drug Reactions in Medical Forums: A Comparison of Algorithms. *ArXiv*.

Natalya F. Noy, Nicholas Griffith, and Mark A. Musen. 2008. Collecting community-based mappings in an ontology repository. In *Proceedings of the 7th International Conference on The Semantic Web*, ISWC '08, page 371–386, Berlin, Heidelberg. Springer-Verlag.

Sameer Pradhan, Noémie Elhadad, Wendy Chapman, Suresh Manandhar, and Guergana Savova. 2014. SemEval-2014 Task 7: Analysis of Clinical Text. In *Proceedings ofthe 8th International Workshop on Semantic Evaluation (SemEval 2014)*, SemEval, pages 54–62.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In *5th Workshop on Energy Efficient Machine Learning and Cognitive Computing - NeurIPS 2019*.

Abeed Sarker, Rachel Ginn, Azadeh Nikfarjam, Karen O'Connor, Karen Smith, Swetha Jayaraman, Tejaswi Upadhaya, and Graciela Gonzalez. 2015. Utilizing social media data for pharmacovigilance: A review. *Journal of Biomedical Informatics*, 54:202–212.

Jingbo Shang, Liyuan Liu, Xiaotao Gu, Xiang Ren, Teng Ren, and Jiawei Han. 2020. Learning named entity tagger using domain-specific dictionary. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*, pages 2054–2064.

Premnath Shenoy and Anand Harugeri. 2015. Elderly patients' participation in clinical trials. *Perspectives in Clinical Research*, 6(4):184.

Yuqi Si, Jingqi Wang, Hua Xu, and Kirk Roberts. 2019. Enhancing clinical concept extraction with contextual embeddings. *Journal of the American Medical Informatics Association*, 26(11):1297–1304.

Leslie N. Smith. 2018. A disciplined approach to neural network hyper-parameters: Part 1 – learning rate, batch size, momentum, and weight decay. *ArXiv*.

Jennifer D ' Souza and Vincent Ng. 2015. Sieve-Based Entity Linking for the Biomedical Domain. In *53rd Annual Meeting ofthe Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 297–302.

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.

Buzhou Tang, Qingcai Chen, Xiaolong Wang, Yonghui Wu, Yaoyun Zhang Phd, Min Jiang, Jingqi Wang, and Hua Xu. 2015. Recognizing Disjoint Clinical Concepts in Clinical Text Using Machine Learning-based Methods. In *AMIA Annu Symp Proc.*, pages 1184–1193.

Buzhou Tang, Jianglu Hu, Xiaolong Wang, and Qingcai Chen. 2018. Recognizing Continuous and Discontinuous Adverse Drug Reaction Mentions from Social Media Using LSTM-CRF. *Wireless Communications and Mobile Computing*, 2018.

Elena Tutubalina, Artur Kadurin, and Zulfat Miftahutdinov. 2020. Fair Evaluation in Concept Normalization : a Large-scale Comparative Analysis for BERT-based Models. In *COLING 2020*.

Özlem Uzuner, Brett R. South, Shuying Shen, and Scott L. DuVall. 2011. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556.

Davy Weissenbacher, Abeed Sarker, Arjun Magge, Ashlynn Daughton, Karen O'Connor, Michael J. Paul, and Graciela Gonzalez-Hernandez. 2019. Overview of the fourth social media mining for health (SMM4H) shared tasks at ACL 2019. In *Proceedings of the Fourth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 21–30, Florence, Italy. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Transformers: State-of-the-art Natural Language Processing. *ArXiv*.

World Health Organisation. 2006. The Safety of Medicines in Public Health Programmes: Pharmacovigilance an essential tool. Technical report, World Health Organisation.

Yaoyun Zhang, Jingqi Wang, Buzhou Tang, Yonghui Wu, Min Jiang, Yukun Chen, and Hua Xu. 2014. UTH_CCB: A report for SemEval 2014 – Task 7 Analysis of Clinical Text. In *Proceedings ofthe 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 802–806.

Maryam Zolnoori, Kin Wah Fung, Timothy B Patrick, Paul Fontelo, Hadi Kharrazi, Anthony Faiola, Nilay D Shah, Yi Shuan Shirley Wu, Christina E Eldredge, Jake Luo, Mike Conway, Jiaxi Zhu, Soo Kyung Park, Kelly Xu, and Hamideh Moayyed. 2019. The PsyTAR dataset: From patients generated narratives to a corpus of adverse drug events and effectiveness of psychiatric medications. *Data in brief*, 24.