



Universiteit  
Leiden  
The Netherlands

## **Knowledge discovery from patient forums: gaining novel medical insights from patient experiences**

Dirkson, A.R.

### **Citation**

Dirkson, A. R. (2022, December 6). *Knowledge discovery from patient forums: gaining novel medical insights from patient experiences*. SIKS Dissertation Series. Retrieved from <https://hdl.handle.net/1887/3492655>

Version: Publisher's Version

License: [Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden](#)

Downloaded from: <https://hdl.handle.net/1887/3492655>

**Note:** To cite this publication please use the final published version (if applicable).

# 7

## FUZZY REPRESENTATION OF DISCONTINUOUS ENTITIES

Edited from: **Anne Dirkson**, Suzan Verberne and Wessel Kraaij (2021), FuzzyBIO: A proposal for Fuzzy Representation of Discontinuous Entities, Proceedings of the 12th Health Text Mining and Information Analysis at EACL 2021.

*For the task of extracting entities (Named Entity Recognition) such as side effects, words in the data are generally represented using the BIO scheme (B-beginning, I-inside, and O-outside) to indicate whether a word is the beginning of an entity, inside an entity or not in an entity. However, this representation does not provide a way of dealing with discontinuous entities.*

*As discontinuous entities occur commonly in the biomedical domain, expansions of the BIO scheme that can handle these entity types are often used (i.e., BIOHD). However, the extra tag types make the NER task more difficult to learn. Therefore, in this chapter, we present FuzzyBIO as an alternative simpler representation scheme in which discontinuous entities are transformed into continuous sequences by annotating the non-entity words in between.*

*We focus on the task of Adverse Drug Event extraction and normalization to compare FuzzyBIO with BIOHD. We find that FuzzyBIO improves the recall of NER for two of the 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 the 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.*

## 7.1. INTRODUCTION

Adverse Drug Events (ADEs), harmful reactions that result from the intake of medication, pose a major health concern [340]. Due to the limitations of clinical trials on the one hand [274] and reporting systems after release on the market on the other hand [130], many ADEs 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 ADEs [171, 262].

One particular challenge for the extraction of ADEs 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. [295] reported that discontinuous mentions in clinical text account for about 10% of all ADE 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)<sup>1</sup> were designed to handle such mentions [241]. Therefore, Tang et al. [295] 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 [151, 275, 353]. Table 7.1 shows examples of concepts represented with the BIOHD scheme.

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

Table 7.1: Examples of discontinuous disjoint (sentence 1) and composite (sentence 2) ADE mentions represented by the BIOHD and FuzzyBIO schemes.

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 chapter 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 ADE 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 7.1.

Aside from improving extraction, using FuzzyBIO instead of BIOHD may also improve subsequent concept normalization, in which ADEs 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 ADEs. As current normalization methods are

<sup>1</sup>This scheme is also called BIOES, BILOU or BMEWO

mostly hampered by errors made during extraction [335], this step may also benefit from simplification of the representation scheme.

Thus, we address two research questions:

- RQ1** To what extent can a fuzzy continuous representation of discontinuous entities improve NER of ADEs? (intrinsic evaluation)
- RQ2** To what extent can this fuzzy representation benefit end-to-end ADE extraction? (extrinsic evaluation)

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

## 7.2. RELATED WORK

The first shared task to deal with discontinuous medical entities was SemEval 2014 Task 7 [241]. Prior to this task, discontinuous entities were often excluded (e.g., Uzuner et al. [315]) or each part was represented as a separate continuous entity and later reassembled [203]. 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 [351].

This scheme was later analyzed 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 [295]. 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. [295] 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<sup>3</sup>, 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. [296] 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 events, this novel representation managed to outperform BIOHD. Similarly, Shang et al. [272] allowed for multiple labels per token for extracting disorders from scientific articles.

Despite its limitations [296], the BIOHD scheme is commonly adopted [151, 275, 353]. We propose an alternative, simpler representation scheme that could improve extraction by being easier to learn.

<sup>2</sup>Code is available at: <https://github.com/AnneDirkson/FuzzyBIO>

<sup>3</sup>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

## 7.3. METHODS

### 7.3.1. THE FUZZYBIO REPRESENTATION SCHEME

As displayed in Table 7.1, FuzzyBIO transforms discontinuous into continuous entities by annotating all tokens in between.<sup>4</sup> Composite entities are combined if they share an entity head. We realize this compresses two separate entities into one (e.g., the entities ‘pain in my hands’ and ‘pain in my upper arms’ in Table 7.1). However, this does not pose a problem to normalization, as the state-of-the-art normalization method [291] includes heuristic rules to split composite entities prior to normalization.

### 7.3.2. NAMED ENTITY RECOGNITION OF ADES

For the NER task itself, we opt for distilBERT (base-cased), a lighter, more computationally efficient version of BERT [260]. We use a one-cycle learning rate (LR) policy [281] 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 [339] with the wrapper ktrain [195] to train our models with the initialization seed set to 1.

### 7.3.3. CONCEPT NORMALIZATION OF ADES

For normalization, we use the state-of-the-art BioSyn method with default parameters [291, 304]. It is possible to provide composite entities as input, as this method splits composite entities prior to normalization using the heuristics by D’ Souza and Ng [70].

Our target ontology is SNOMED-CT<sup>5</sup>. 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<sup>6</sup>, before training the normalization model. If there is a direct mapping in the community based mappings in BioPortal [220] 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 [220] using the REST API and from the UMLS using pymedtermino [170].

### 7.3.4. 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. [194], an entity is considered a true positive if any part of the annotated adverse drug event is correctly identified (i.e., overlaps with the predicted ADE 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.

<sup>4</sup>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.

<sup>5</sup><https://www.nlm.nih.gov/healthit/snomedct/index.html>

<sup>6</sup>[https://www.nlm.nih.gov/research/umls/Snomed/core\\_subset.html](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 7.2: Size of data sets.

## 7.4. DATA

We use two data sets of social media posts annotated for ADEs: CADEC [151] and PsyTAR [353]. The former was also used by Tang et al. [296]. 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 [241] that builds on the CLEF eHealth 2013 corpus used by Tang et al. [295]. See Table 7.2 for more details. Data sets were split into 10 folds stratified on the presence of an ADE. 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.

## 7.5. RESULTS

Data	Scheme	Token level			Entity level recall		
		Micro F <sub>1</sub>	P	R	Continuous	Disjoint	Composite
CADEC	BIOHD	0.586	<b>0.636</b>	0.555	42.0%	55.4%	64.6%
	FuzzyBIO	<b>0.596</b>	0.612	<b>0.584</b>	<b>43.1%</b>	<b>58.9%</b>	<b>65.2%</b>
PsyTAR	BIOHD	<b>0.771</b>	<b>0.751</b>	<b>0.797</b>	<b>87.4%</b>	83.8%	79.5%
	FuzzyBIO	0.762	0.747	0.780	84.9%	<b>84.4%</b>	<b>87.5%</b>
CLEF	BIOHD	<b>0.312</b>	<b>0.286</b>	0.345	35.6%	52.8%	69.1%
	FuzzyBIO	0.309	0.276	<b>0.352</b>	<b>36.2%</b>	<b>55.2%</b>	<b>76.6%</b>

Table 7.3: Intrinsic evaluation of NER. Results are the average of a 10 fold CV.

### 7.5.1. INTRINSIC EVALUATION

As can be seen in Table 7.3, 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 and +3.6 for disjoint and +0.6 and +0.7 for composite entities). For the remaining data set (PsyTAR), the 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.

### 7.5.2. EXTRINSIC EVALUATION

As can be seen in Table 7.4, 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%	<b>35.9%</b>	21.2%
	FuzzyBIO	<b>24.3%</b>	20.4%	<b>22.5%</b>
PsyTAR	BIOHD	<b>43.6%</b>	<b>26.1%</b>	<b>10.6%</b>
	FuzzyBIO	42.8%	25.0%	7.5%
CLEF	BIOHD	21.7%	<b>25.8%</b>	26.5%
	FuzzyBIO	<b>22.1%</b>	4.8%	<b>32.2%</b>

Table 7.4: Extrinsic evaluation of ADE extraction. Results are the average of a 10 fold CV.

## 7.6. 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 instead of 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 ADE 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 the 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 comparable 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 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.

## 7.7. CONCLUSION

We expect FuzzyBIO to be most beneficial for 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 experimentally comparing which of the two is more effective for any data set that includes discontinuous entities.



