An Effective Transition-based Model for Discontinuous NER

Xiang Dai[1,2]   **Sarvnaz Karimi**[1]   **Ben Hachey**[3]   **Cecile Paris**[1]

[1]CSIRO Data61, Sydney, Australia
[2]University of Sydney, Sydney, Australia
[3]Harrison.ai, Sydney, Australia
`{dai.dai,sarvnaz.karimi,cecile.paris}@csiro.au`
`ben.hachey@gmail.com`

Abstract

Unlike widely used Named Entity Recognition (NER) data sets in generic domains, biomedical NER data sets often contain mentions consisting of discontinuous spans. Conventional sequence tagging techniques encode Markov assumptions that are efficient but preclude recovery of these mentions. We propose a simple, effective transition-based model with generic neural encoding for discontinuous NER. Through extensive experiments on three biomedical data sets, we show that our model can effectively recognize discontinuous mentions without sacrificing the accuracy on continuous mentions.

1 Introduction

Named Entity Recognition (NER) is a critical component of biomedical natural language processing applications. In pharmacovigilance, it can be used to identify adverse drug events in consumer reviews in online medication forums, alerting medication developers, regulators and clinicians (Leaman et al., 2010; Sarker et al., 2015; Karimi et al., 2015b). In clinical settings, NER can be used to extract and summarize key information from electronic medical records such as conditions hidden in unstructured doctors' notes (Feblowitz et al., 2011; Wang et al., 2018b). These applications require identification of complex mentions not seen in generic domains (Dai, 2018).

Widely used sequence tagging techniques (*flat model*) encode two assumptions that do not always hold: (1) mentions do not nest or overlap, therefore each token can belong to at most one mention; and, (2) mentions comprise continuous sequences of tokens. Nested entity recognition addresses violations of the first assumption (Lu and Roth, 2015; Katiyar and Cardie, 2018; Sohrab and Miwa, 2018; Ringland et al., 2019). However, the violation of

The left atrium is mildly dilated .
E1   E1

have much muscle pain and fatigue .
E2
E3   E3

Figure 1: Examples involving discontinuous mentions, taken from the ShARe 13 (Pradhan et al., 2013) and CADEC (Karimi et al., 2015a) data sets, respectively. The first example contains a discontinuous mention *'left atrium dilated'*, the second example contains two mentions that overlap: *'muscle pain'* and *'muscle fatigue'* (discontinuous).

the second assumption is comparatively less studied and requires handling discontinuous mentions (see examples in Figure 1).

In contrast to continuous mentions which are often short spans of text, *discontinuous mentions* consist of **components** that are separated by **intervals**. Recognizing discontinuous mentions is particularly challenging as exhaustive enumeration of possible mentions, including discontinuous and overlapping spans, is exponential in sentence length. Existing approaches for discontinuous NER either suffer from high time complexity (McDonald et al., 2005) or ambiguity in translating intermediate representations into mentions (Tang et al., 2013a; Metke-Jimenez and Karimi, 2016; Muis and Lu, 2016). In addition, current art uses traditional approaches that rely on manually designed features, which are tailored to recognize specific entity types. Also, these features usually do not generalize well in different genres (Leaman et al., 2015).

Motivations   The main motivation for recognizing discontinuous mentions is that they usually represent *compositional concepts* that differ from concepts represented by individual components. For example, the mention *'left atrium dilated'* in the first example of Figure 1 describes a disorder which has its own CUI (Concept Unique Identi-

5860

fier) in UMLS (Unified Medical Language System), whereas both *'left atrium'* and *'dilated'* also have their own CUIs. We argue that, in downstream applications such as pharmacovigilance and summarization, recognizing these discontinuous mentions that refer to disorders or symptoms is more useful than recognizing separate components which may refer to body locations or general feelings.

Another important characteristic of discontinuous mentions is that they usually *overlap*. That is, several mentions may share components that refer to the same body location (e.g., *'muscle'* in *'muscle pain and fatigue'*), or the same feeling (e.g., *'Pain'* in *'Pain in knee and foot'*). Separating these overlapping mentions rather than identifying them as a single mention is important for downstream tasks, such as entity linking where the assumption is that the input mention refers to one entity (Shen et al., 2015).

Contributions We propose an end-to-end transition-based model with generic neural encoding that allows us to leverage specialized actions and attention mechanism to determine whether a span is the component of a discontinuous mention or not.[1] We evaluate our model on three biomedical data sets with a substantial number of discontinuous mentions and demonstrate that our model can effectively recognize discontinuous mentions without sacrificing the accuracy on continuous mentions.

2 Prior Work

Existing methods on discontinuous NER can be mainly categorized into two categories: token level approach, based on sequence tagging techniques, and sentence level approach, where a combination of mentions within a sentence is jointly predicted (Dai, 2018).

Token level approach Sequence tagging model takes a sequence of tokens as input and outputs a tag for each token, composed of a position indicator (e.g., BIO schema) and an entity type. The vanilla BIO schema cannot effectively represent discontinuous, overlapping mentions, therefore, some studies overcome this limitation via expanding the BIO tag set (Tang et al., 2013a; Metke-Jimenez and Karimi, 2016; Dai et al., 2017; Tang et al., 2018). In addition to BIO indicators, four new position indicators are introduced in (Metke-Jimenez and

Karimi, 2016) to represent discontinuous mentions that may overlap:

- **BH**: **B**eginning of **H**ead, defined as the components shared by multiple mentions;

- **IH**: **I**ntermediate of **H**ead;

- **BD**: **B**eginning of **D**iscontinuous body, defined as the exclusive components of a discontinuous mention; and

- **ID**: **I**ntermediate of **D**iscontinuous body.

Sentence level approach Instead of predicting whether each token belongs to an entity mention and its role in the mention, sentence level approach predicts a combination of mentions within a sentence. A hypergraph, proposed by Lu and Roth (2015) and extended in (Muis and Lu, 2016), can compactly represent discontinuous and overlapping mentions in one sentence. A sub-hypergraph of the complete hypergraph can, therefore, be used to represent a combination of mentions in the sentence. For the token at each position, there can be six different node types:

- **A**: mentions that start from the current token or a future token;

- **E**: mentions that start from the current token;

- **T**: mentions of a certain entity type that start from the current token;

- **B**: mentions that contain the current token;

- **O**: mentions that have an interval at the current token;

- **X**: mentions that end at the current token.

Using this representation, a single entity mention can be represented as a path from node **A** to node **X**, incorporating at least one node of type **B**.

Note that both token level and sentence level approaches predict first an intermediate representation of mentions (e.g., a sequence of tags in (Metke-Jimenez and Karimi, 2016) and a sub-hypergraph in (Muis and Lu, 2016)), which are then decoded into the final mentions. During the final decoding stage, both models suffer from some level of ambiguity. Taking the sequence tagging model using BIO variant schema as an example, even if the model can correctly predict the gold sequence of tags for the example sentence 'muscle pain and

---

[1]Code available at GitHub: https://bit.ly/2XazEAO

fatigue' (BH I O BD), it is still not clear whether the token *'muscle'* forms a mention by itself, because the same sentence containing three mentions (*'muscle'*, *'muscle pain'* and *'muscle fatigue'*) can be encoded using the same gold sequence of tags. We refer to a survey by (Dai, 2018) for more discussions on these models, and (Muis and Lu, 2016) for a theoretical analysis of ambiguity of these models.

Similar to prior work, our proposed transition-based model uses an intermediate representation (i.e., a sequence of actions). However, it does not suffer from this ambiguity issue. That is, the output sequence of actions can always be unambiguously decoded into mention outputs.

The other two methods that focus on the discontinuous NER problem in literature are described in (McDonald et al., 2005; Wang and Lu, 2019). McDonald et al. (2005) solve the NER task as a structured multi-label classification problem. Instead of starting and ending indices, they represent each entity mention using the set of token positions that belong to the mention. This representation is flexible, as it allows mentions consisting of discontinuous tokens and does not require mentions to exclude each other. However, this method suffers from high time complexity. Tang et al. (2018) compare this representation with BIO variant schema proposed in (Metke-Jimenez and Karimi, 2016), and found that they achieve competitive F_1 scores, although the latter method is more efficient. A two-stage approach that first detects all components and then combines components into discontinuous mentions based on a classifier's decision was explored in recent work by Wang and Lu (2019).

Discontinuous NER vs. Nested NER   Although discontinuous mentions may overlap, we discriminate this overlapping from the one in nested NER. That is, if one mention is completely contained by the other, we call mentions involved nested entity mentions. In contrast, overlapping in discontinuous NER is usually that two mentions overlap, but no one is completely contained by the other. Most of existing nested NER models are built to tackle the complete containing structure (Finkel and Manning, 2009; Lu and Roth, 2015), and they cannot be directly used to identify overlapping mentions studied in this paper, nor mention the discontinuous mentions. However, we note that there is a possible perspective to solve discontinuous NER task by adding fine-grained entity types into the schema. Taking the second sentence in Figure 1
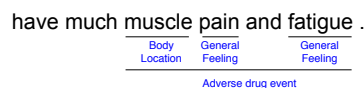

Figure 2: Examples involving Nested mentions.

as an example, we can add two new entity types: 'Body Location' and 'General Feeling', and then annotate 'muscle pain and fatigue' as a 'Adverse drug event' mention, 'muscle' as a 'Body Location' mention, and 'pain' and 'fatigue' as 'General Feeling' mentions (Figure 2). Then the discontinuous NER task can be converted into a Nested NER task.

3   Model

Transition-based models, due to their high efficiency, are widely used for NLP tasks, such as parsing and entity recognition (Chen and Manning, 2014; Lample et al., 2016; Lou et al., 2017; Wang et al., 2018a). The model we propose for discontinuous NER is based on the shift-reduce parser (Watanabe and Sumita, 2015; Lample et al., 2016) that employs a **stack** to store partially processed spans and a **buffer** to store unprocessed tokens. The learning problem is then framed as: given the state of the parser, predict an action which is applied to change the state of the parser. This process is repeated until the parser reaches the end state (i.e., the stack and buffer are both empty).

The main difference between our model and the ones in (Watanabe and Sumita, 2015; Lample et al., 2016) is the set of transition actions. Watanabe and Sumita (2015) use SHIFT, REDUCE, UNARY, FINISH, and IDEA for the constituent parsing system. Lample et al. (2016) use SHIFT, REDUCE, OUT for the flat NER system. Inspired by these models, we design a set of actions specifically for recognizing discontinuous and overlapping structure. There are in total six actions in our model:

- SHIFT moves the first token from the buffer to the stack; it implies this token is part of an entity mention.

- OUT pops the first token of the buffer, indicating it does not belong to any mention.

- COMPLETE pops the top span of the stack, outputting it as an entity mention. If we are interested in multiple entity types, we can extend this action to COMPLETE-y which labels the mention with entity type y.

5862

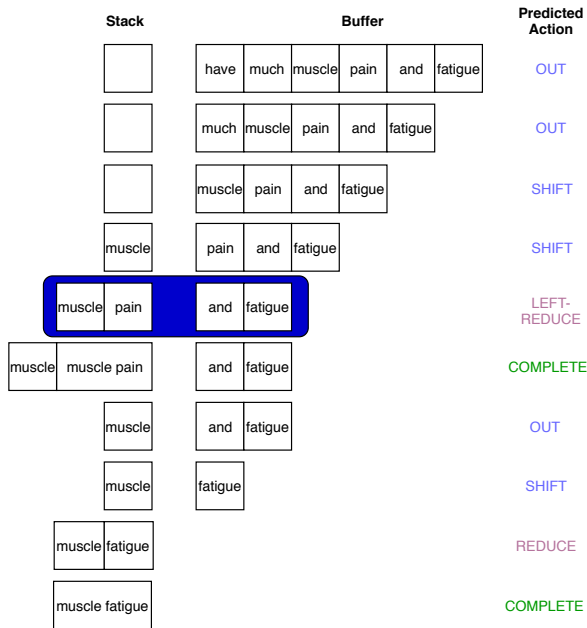| Stack	Buffer	Predicted Action
	have | much | muscle | pain | and | fatigue	OUT
	much | muscle | pain | and | fatigue	OUT
	muscle | pain | and | fatigue	SHIFT
muscle	pain | and | fatigue	SHIFT
muscle | **pain**	**and** | fatigue	LEFT-REDUCE
muscle | muscle pain	and | fatigue	COMPLETE
muscle	and | fatigue	OUT
muscle	fatigue	SHIFT
muscle | fatigue		REDUCE
muscle fatigue		COMPLETE

Figure 3: An example sequence of transitions. Given the states of stack and buffer (blue highlighted), as well as the previous actions, predict the next action (i.e., LEFT-REDUCE) which is then applied to change the states of stack and buffer.

- REDUCE pops the top two spans s_0 and s_1 from the stack and concatenates them as a new span which is then pushed back to the stack.

- LEFT-REDUCE is similar to the REDUCE action, except that the span s_1 is kept in the stack. This action indicates the span s_1 is involved in multiple mentions. In other words, several mentions share s_1 which could be a single token or several tokens.

- RIGHT-REDUCE is the same as LEFT-REDUCE, except that s_0 is kept in the stack.

Figure 3 shows an example about how the parser recognizes entity mentions from a sentence. Note that, given one parser state, not all types of actions are valid. For example, if the stack does not contain any span, only SHIFT and OUT actions are valid because all other actions involve popping spans from the stack. We employ hard constraints that we only select the most likely action from valid actions.

3.1 Representation of the Parser State

Given a sequence of N tokens, we first run a bi-directional LSTM (Graves et al., 2013) to derive the contextual representation of each token. Specif-

ically, for the i-th token in the sequence, its representation can be denoted as:

$$\tilde{\mathbf{c}}_{\mathbf{i}} = \left[\overrightarrow{\text{LSTM}}(\mathbf{t_0}, \dots, \mathbf{t_i}); \overleftarrow{\text{LSTM}}(\mathbf{t_i}, \dots, \mathbf{t_{N-1}}) \right],$$

where $\mathbf{t_i}$ is the concatenation of the embeddings for the i-th token, its character level representation learned using a CNN network (Ma and Hovy, 2016). Pretrained contextual word representations have shown its usefulness on improving various NLP tasks. Here, we can also concatenate pretrained contextual word representations using ELMo (Peters et al., 2018) with $\tilde{\mathbf{c}}_{\mathbf{i}}$, resulting in:

$$\mathbf{c_i} = [\tilde{\mathbf{c}}_{\mathbf{i}}; \text{ELMo}_i], \qquad (1)$$

where ELMo_i is the output representation of pre-trained ELMo models (frozen) for the i-th token. These token representations $\mathbf{c}$ are directly used to represent tokens in the buffer. We also explore a variant that uses the output of pretrained BERT (Devlin et al., 2019) as token representations $\mathbf{c}$, and fine-tune the BERT model. However, this fine-tuning approach with BERT does not achieve as good performance as feature extraction approach with ELMo (Peters et al., 2019).

Following the work in (Dyer et al., 2015), we use Stack-LSTM to represent spans in the stack. That is, if a token is moved from the buffer to the stack, its representation is learned using:

$$\mathbf{s_0} = \text{Stack-LSTM}(\mathbf{s_D} \dots \mathbf{s_1}; \mathbf{c}_{\text{SHIFT}}),$$

where D is the number of spans in the stack. Once REDUCE related actions are applied, we use a multi-layer perceptron to learn the representation of the concatenated span. For example, the REDUCE action takes the representation of the top two spans in the stack: $\mathbf{s_0}$ and $\mathbf{s_1}$, and produces a new span representation:

$$\tilde{\mathbf{s}} = \mathbf{W}^T[\mathbf{s_0}; \mathbf{s_1}] + b,$$

where $\mathbf{W}$ and b denote the parameters for the composition function. The new span representation $\tilde{\mathbf{s}}$ is pushed back to the stack to replace the original two spans: $\mathbf{s_0}$ and $\mathbf{s_1}$.

3.2 Capturing Discontinuous Dependencies

We hypothesize that the interactions between spans in the stack and tokens in the buffer are important factors in recognizing discontinuous mentions. Considering the example in Figure 3, a span in the

stack (e.g., *'muscle'*) may need to combine with a future token in the buffer (e.g., *'fatigue'*). To capture this interaction, we use multiplicative attention (Luong et al., 2015) to let the span in the stack s_i learn which token in the buffer to attend, and thus a weighted sum of the representation of tokens in the buffer B:

$$s_i^a = \text{softmax}(s_i^T W_i^a B)B. \qquad (2)$$

We use distinct W_i^a for s_i separately.

3.3 Selecting an Action

Finally, we build the parser representation as the concatenation of the representation of top three spans from the stack (s_0, s_1, s_2) and its attended representation (s_0^a, s_1^a, s_2^a), as well as the representation of the previous action a, which is learned using a simple unidirectional LSTM. If there are less than 3 spans in the stack or no previous action, we use randomly initialized vectors s_{empty} or a_{empty} to replace the corresponding vector. This parser representation is used as input for the final softmax prediction layer to select the next action.

4 Data sets

Although some text annotation tools, such as BRAT (Stenetorp et al., 2012), allow discontinuous annotations, corpora annotated with a large number of discontinuous mentions are still rare. We use three data sets from the biomedical domain: CADEC (Karimi et al., 2015a), ShARe 13 (Pradhan et al., 2013) and ShARe 14 (Mowery et al., 2014). Around 10% of mentions in these three data sets are discontinuous. The descriptive statistics are listed in Table 1.

CADEC is sourced from *AskaPatient*[2], a forum where patients can discuss their experiences with medications. The entity types in CADEC include drug, Adverse Drug Event (ADE), disease and symptom. We only use ADE annotations because only the ADEs involve discontinuous annotations. This also allows us to compare our results directly against previously reported results (Metke-Jimenez and Karimi, 2016; Tang et al., 2018). ShARe 13 and 14 focus on the identification of disorder mentions in clinical notes, including discharge summaries, electrocardiogram, echocardiogram, and radiology reports (Johnson et al., 2016). A disorder mention is defined as any span of text which can be

	CADEC	ShARe 13	ShARe 14
Text type	online posts	clinical notes	clinical notes
Entity type	ADE	Disorder	Disorder
# Documents	1,250	298	433
# Tokens	121K	264K	494K
# Sentences	7,597	18,767	34,618
# Mentions	6,318	11,161	19,131
# Disc.M	675 (10.6)	1,090 (9.7)	1,710 (8.9)
Avg mention L.	2.7	1.8	1.7
Avg Disc.M L.	3.5	2.6	2.5
Avg interval L.	3.3	3.0	3.2
Discontinuous Mentions			
2 components	650 (95.7)	1,026 (94.3)	1,574 (95.3)
3 components	27 (3.9)	62 (5.6)	76 (4.6)
4 components	2 (0.2)	0 (0.0)	0 (0.0)
No overlap	82 (12.0)	582 (53.4)	820 (49.6)
Overlap at left	351 (51.6)	376 (34.5)	616 (37.3)
Overlap at right	152 (22.3)	102 (9.3)	170 (10.3)
Multiple overlaps	94 (13.8)	28 (2.5)	44 (2.6)
Continuous Mentions			
Overlap	326 (5.7)	157 (1.5)	228 (1.3)

Table 1: The descriptive statistics of the data sets. ADE: adverse drug events; Disc.M: discontinuous mentions; Disc.M L.: discontinuous mention length, where intervals are not counted. Numbers in parentheses are the percentage of each category.

mapped to a concept in the disorder semantic group of SNOMED-CT (Cornet and de Keizer, 2008).

Although these three data sets share similar field (the subject matter of the content being discussed), the tenor (the participants in the discourse, their relationships to each other, and their purposes) of CADEC is very different from the ShARe data sets (Dai et al., 2019). In general, laymen (i.e., in CADEC) tend to use idioms to describe their feelings, whereas professional practitioners (i.e., in ShARe) tend to use compact terms for efficient communications. This also results in different features of discontinuous mentions between these data sets, which we will discuss further in § 7.

Experimental Setup As CADEC does not have an official train-test split, we follow Metke-Jimenez and Karimi (2016) and randomly assign 70% of the posts as the training set, 15% as the development set, and the remaining posts as the test set. [3] The train-test splits of ShARe 13 and 14 are both from their corresponding shared task settings, except that we randomly select 10% of documents from each training set as the development set. Micro

---

[2]https://www.askapatient.com/

[3]These splits can be downloaded from https://bit.ly/2XazEAO.

average strict match F_1 score is used to evaluate the effectiveness of the model. The trained model which is most effective on the development set, measured using the F_1 score, is used to evaluate the test set.

5 Baseline Models

We choose one flat NER model which is strong at recognizing continuous mentions, and two discontinuous NER models as our baseline models:

Flat model To train the flat model on our data sets, we use an off-the-shelf framework: Flair (Akbik et al., 2018), which achieves the state-of-the-art performance on CoNLL 03 data set. Recall that the flat model cannot be directly applied to data sets containing discontinuous mentions. Following the practice in (Stanovsky et al., 2017), we replace the discontinuous mention with the shortest span that fully covers it, and merge overlapping mentions into a single mention that covers both. Note that, different from (Stanovsky et al., 2017), we apply these changes only on the training set, but not on the development set and the test set.

BIO extension model The original implementation in (Metke-Jimenez and Karimi, 2016) used a CRF model with manually designed features. We report their results on CADEC in Table 2 and re-implement a BiLSTM-CRF-ELMo model using their tag schema (denoted as 'BIO Extension' in Table 2).

Graph-based model The original paper of (Muis and Lu, 2016) only reported the evaluation results on sentences which contain at least one discontinuous mention. We use their implementation to train the model and report evaluation results on the whole test set (denoted as 'Graph' in Table 2). We argue that it is important to see how a discontinuous NER model works not only on the discontinuous mentions but also on all the mentions, especially since, in real data sets, the ratio of discontinuous mentions cannot be made a priori.

We do not choose the model proposed in (Wang and Lu, 2019) as the baseline model, because it is based on a strong assumption about the ratio of discontinuous mentions. Wang and Lu (2019) train and evaluate their model on sentences that contain at least one discontinuous mention. Our early experiments show that the effectiveness of their model strongly depends on this assumption.

In contrast, we train and evaluate our model in a more practical setting where the number of continuous mentions is much larger than the one of discontinuous mentions.

6 Experimental Results

When evaluated on the whole test set, our model outperforms three baseline models, as well as over previous reported results in the literature, in terms of recall and F_1 scores (Table 2).

The graph-based model achieves highest precision, but with substantially lower recall, therefore obtaining lowest F_1 scores. In contrast, our model improves recall over flat and BIO extension models as well as previously reported results, without sacrificing precision. This results in more balanced precision and recall. Improved recall is especially encouraging for our motivating pharmacovigilance and medical record summarization applications, where recall is at least as important as precision.

Effectiveness on recognizing discontinuous mentions Recall that only 10% of mentions in these three data sets are discontinuous. To evaluate the effectiveness of our proposed model on recognizing discontinuous mentions, we follow the evaluation approach in (Muis and Lu, 2016) where we construct a subset of test set where only sentences with at least one discontinuous mention are included (Left part of Table 3). We also report the evaluation results when only discontinuous mentions are considered (Right part of Table 3). Note that sentences in the former setting usually contain continuous mentions as well, including those involved in overlapping structure (e.g., 'muscle pain' in the sentence 'muscle pain and fatigue'). Therefore, the flat model, which cannot predict any discontinuous mentions, still achieves 38% F_1 on average when evaluated on these sentences with at least one discontinuous mention, but 0% F_1 when evaluated on discontinuous mentions only.

Our model again achieves the highest F_1 and recall in all three data sets under both settings. The comparison between these two evaluation results also shows the necessity of comprehensive evaluation settings. The BIO E. model outperforms the graph-based model in terms of F_1 score on CADEC, when evaluated on sentences with discontinuous mentions. However, it achieves only 1.8 F_1 when evaluated on discontinuous mentions only. The main reason is that most of discontinuous mentions in CADEC are involved in overlapping

Model	CADEC			ShARe 13			ShARe 14		
	P	R	F	P	R	F	P	R	F
(Metke-Jimenez and Karimi, 2016)	64.4	56.5	60.2	–	–	–	–	–	–
(Tang et al., 2018)	67.8	64.9	66.3	–	–	–	–	–	–
(Tang et al., 2013b)	–	–	–	80.0	70.6	75.0	–	–	–
Flat	65.3	58.5	61.8	78.5	66.6	72.0	76.2	76.7	76.5
BIO Extension	68.7	66.1	67.4	77.0	72.9	74.9	74.9	78.5	76.6
Graph	**72.1**	48.4	58.0	**83.9**	60.4	70.3	**79.1**	70.7	74.7
Ours	68.9	69.0	69.0	80.5	**75.0**	**77.7**	78.1	**81.2**	**79.6**

Table 2: Evaluation results on the whole test set in terms of precision, recall and F_1 score. The original ShARe 14 task focuses on template filling of disorder attributes: that is, given a disorder mention, recognize the attribute from its context. In this work, we use its mention annotations and frame the task as a discontinuous NER task.

| | Sentences with discontinuous mentions | | | | | | | | | Discontinuous mentions only | | | | | | | | |
| | CADEC | | | ShARe 13 | | | ShARe 14 | | | CADEC | | | ShARe 13 | | | ShARe 14 | | |
Model	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
Flat	50.2	36.7	42.4	43.5	28.1	34.2	41.5	31.9	36.0	0	0	0	0	0	0	0	0	0
BIO E.	63.8	52.0	57.3	51.8	39.5	44.8	37.5	38.4	37.9	5.8	1.0	1.8	39.7	12.3	18.8	8.8	4.5	6.0
Graph	**69.5**	43.2	53.3	**82.3**	47.4	60.2	60.0	52.8	56.2	**60.8**	14.8	23.9	78.4	36.6	50.0	42.7	39.5	41.1
Ours	66.5	**64.3**	**65.4**	70.5	**56.8**	**62.9**	61.9	**64.5**	**63.1**	41.2	**35.1**	**37.9**	78.5	39.4	52.5	56.1	43.8	49.2

Table 3: Evaluation results on sentences that contain at least one discontinuous mention (left part) and on discontinuous mentions only (right part).

structure (88%, cf. Table 1), and the BIO E. model is better than the graph-based model at recognizing these continuous mentions. On ShARe 13 and 14, where the portion of discontinuous mentions involved in overlapping is much less than on CADEC, the graph-based model clearly outperforms BIO E. model in both evaluation settings.

7 Analysis

We start our analysis from characterizing discontinuous mentions from the three data sets. Then we measure the behaviors of our model and two discontinuous NER models on the development sets based on characteristics identified and attempt to draw conclusions from these measurements.

7.1 Characteristics of Discontinuous Mentions

Recall that discontinuous mentions usually represent compositional concepts that consist of multiple components. Therefore, discontinuous mentions are usually longer than continuous mentions (Table 1). In addition, intervals between components make the total length of span involved even longer. Previous work shows that flat NER performance degrades when applied on long mentions (Augenstein et al., 2017; Xu et al., 2017).

Another characteristic of discontinuous men-

tions is that they usually *overlap* (cf. § 1). From this perspective, we can categorize discontinuous mentions into four categories:

- No overlap: in such cases, the discontinuous mention can be intervened by severity indicators (e.g., *'is mildly'* in sentence *'left atrium is mildly dilated'*), preposition (e.g., *'on my'* in sentence *'...rough on my stomach...'*) and so on. This category accounts for half of discontinuous mentions in the ShARe data sets but only 12% in CADEC (Table 1).

- Left overlap: the discontinuous mention shares one component with other mentions, and the shared component is at the beginning of the discontinuous mention. This is usually accompanied with coordination structure (e.g., the shared component *'muscle'* in *'muscle pain and fatigue'*). Conjunctions (e.g., *'and'*, *'or'*) are clear indicators of the coordination structure. However, clinical notes are usually written by practitioners under time pressure. They often use commas or slashes rather than conjunctions. This category accounts for more than half of discontinuous mentions in CADEC and one third in ShARe.

- Right overlap: similar to left overlap, although the shared component is at the end. For ex-

(a) CADEC   (b) ShARe 13   (c) ShARe 14
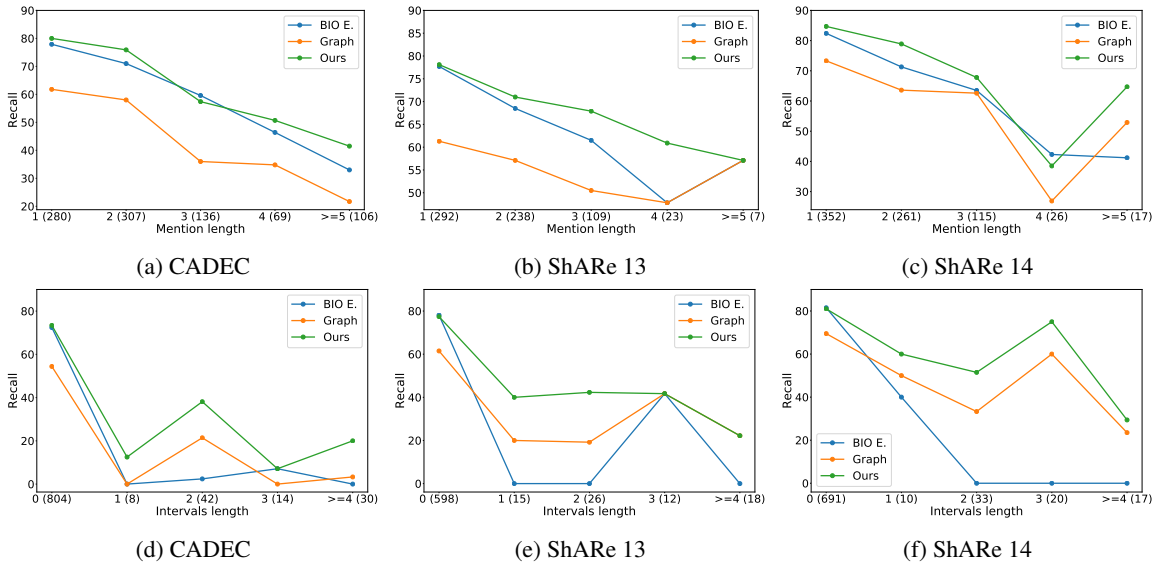
(d) CADEC   (e) ShARe 13   (f) ShARe 14

Figure 4: The impact of mention length and interval length on recall. Mentions with interval length of zero are continuous mentions. Numbers in parentheses are the number of gold mentions.

ample, *'hip/leg/foot pain'* contains three mentions that share *'pain'*.

- Multi-overlap: the discontinuous mention shares multiple components with the others, which usually forms *crossing compositions*. For example, the sentence *'Joint and Muscle Pain / Stiffness'* contains four mentions: *'Joint Pain'*, *'Joint Stiffness'*, *'Muscle Stiffness'* and *'Muscle Pain'*, where each discontinuous mention share two components with the others.

7.2 Impact of Overlapping Structure

Previous study shows that the intervals between components can be problematic for coordination boundary detection (Ficler and Goldberg, 2016). Conversely, we want to observe whether the overlapping structure may help or hinder discontinuous entity recognition. We categorize discontinuous mentions into different subsets, described in § 7.1, and measure the effectiveness of different discontinuous NER models on each category.

From Table 4, we find that our model achieves better results on discontinuous mentions belonging to 'No overlap' category on ShARe 13 and 14, and 'Left overlap' category on CADEC and ShARe 14. Note that 'No overlap' category accounts for half of discontinuous mentions in ShARe 13 and 14, whereas 'Left overlap' accounts for half in CADEC (Table 1). Graph-based model achieves better results on 'Right overlap' category. On the 'Multi-overlap' category, no models is effective, which

emphasizes the challenges of dealing with this syntactic phenomena. We note, however, the portion of discontinuous mentions belonging to this category is very small in all three data sets.

Although our model achieves better results on 'No overlap' category on ShARe 13 and 14, it does not predict correctly any discontinuous mention belonging to this category on CADEC. The ineffectiveness of our model, as well as other discontinuous NER models, on CADEC 'No overlap' category can be attributed to two reasons: 1) the number of discontinuous mentions belonging to this category in CADEC is small (around 12%), rending the learning process more difficult. 2) the gold annotations belonging to this category are inconsistent from a linguistic perspective. For example, severity indicators are annotated as the interval of the discontinuous mention sometimes, but not often. Note that this may be reasonable from a medical perspective, as some symptoms are roughly grouped together no matter their severity, whereas some symptoms are linked to different concepts based on their severity.

7.3 Impact of Mention and Interval Length

We conduct experiments to measure the ability of different models on recalling mentions of different lengths, and to observe the impact of interval lengths. We found that the recall of all models decreases with the increase of mention length in general (Figure 4 (a – c)), which is similar to previous observations in the literature on flat men-

	Model	CADEC		ShARe 13		ShARe 14	
		#	F	#	F	#	F
No ⦾	BIO E.		0.0		7.5		0.0
	Graph	9	0.0	41	32.1	39	45.2
	Ours		0.0		**36.1**		**57.1**
Left ⦾	BIO E.		6.0		25.0		15.7
	Graph	54	9.2	11	**45.5**	30	37.7
	Ours		**28.6**		33.3		**49.2**
Right ⦾	BIO E.		0.0		0.0		0.0
	Graph	16	**45.2**	19	**21.4**	5	0.0
	Ours		29.3		13.3		0.0
Multi ⦾	BIO E.		0.0		–		0.0
	Graph	15	0.0	0	–	6	0.0
	Ours		0.0		–		0.0

Table 4: Evaluation results on different categories of discontinuous mentions. '#' columns show the number of gold discontinuous mentions in development set of each category. ⦾: overlap.

tions. However, the impact of interval length is not straightforward. Mentions with very short interval lengths are as difficult as those with very long interval lengths to be recognized (Figure 4 (d – f)). On CADEC, discontinuous mentions with interval length of 2 are easiest to be recognized (Figure 4 (d)), whereas those with interval length of 3 are easiest on ShARe 13 and 14. We hypothesize this also relates to annotation inconsistency, because very short intervals may be overlooked by annotators.

In terms of model comparison, our model achieves highest recall in most settings. This demonstrates our model is effective to recognize both continuous and discontinuous mentions with various lengths. In contrast, the BIO E. model is only strong at recalling continuous mentions (outperforming the graph-based model), but fails on discontinuous mentions (interval lengths > 0).

7.4 Example Predictions

We find that previous models often fail to identify discontinuous mentions that involve long and overlapping spans. For example, the sentence 'Severe joint pain in the shoulders and knees.' contains two mentions: 'Severe joint pain in the shoulders' and 'Severe joint pain in the knees'. Graph-based model does not identify any mention from this sentence, resulting in a low recall. The BIO extension model predicts most of these tags (8 out of 9) correctly, but fails to decode into correct mentions (predict 'Severe joint pain in the', resulting in a false positive, while it misses 'Severe joint pain in the shoulders'). In contrast, our model correctly identifies both of these two mentions.

No model can fully recognize mentions which form crossing compositions. For example, the sentence 'Joint and Muscle Pain / Stiffness' contains four mentions: 'Joint Pain', 'Joint Stiffness', 'Muscle Stiffness' and 'Muscle Pain', all of which share multiple components with the others. Our model correctly predicts 'Joint Pain' and 'Muscle Pain', but it mistakenly predicts 'Stiffness' itself as a mention.

8 Summary

We propose a simple, effective transition-based model that can recognize discontinuous mentions without sacrificing the accuracy on continuous mentions. We evaluate our model on three biomedical data sets with a substantial number of discontinuous mentions. Comparing against two existing discontinuous NER models, our model is more effective, especially in terms of recall.

References

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *COLING*, pages 1638–1649, Santa Fe, New Mexico.

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 *SemEval*, pages 546–555, Vancouver, Canada.

Danqi Chen and Christopher Manning. 2014. A fast and accurate dependency parser using neural networks. In *EMNLP*, pages 740–750, Doha, Qatar.

Ronald Cornet and Nicolette de Keizer. 2008. Forty years of SNOMED: a literature review. In *BMC Med Inform Decis Mak*, volume 8, page S2.

Xiang Dai. 2018. Recognizing complex entity mentions: A review and future directions. In *ACL@SRW*, pages 37–44, Melbourne, Australia.

Xiang Dai, Sarvnaz Karimi, Ben Hachey, and Cecile Paris. 2019. Using similarity measures to select pretraining data for ner. In *NAACL*, pages 1460–1470, Minneapolis, Minnesota.

Xiang Dai, Sarvnaz Karimi, and Cècile Paris. 2017. Medication and adverse event extraction from noisy text. In *ALTA*, pages 79–87, Brisbane, Australia.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pages 4171–4186, Minneapolis, Minnesota.

Chris Dyer, Miguel Ballesteros, Wang Ling, Austin Matthews, and Noah A. Smith. 2015. Transition-based dependency parsing with stack long short-term memory. In *ACL-IJCNLP*, pages 334–343, Beijing, China.

Joshua C Feblowitz, Adam Wright, Hardeep Singh, Lipika Samal, and Dean F Sittig. 2011. Summarization of clinical information: a conceptual model. *J Biomed Inform*, 44(4):688–699.

Jessica Ficler and Yoav Goldberg. 2016. A neural network for coordination boundary prediction. In *EMNLP*, pages 23–32, Austin, Texas.

Jenny Rose Finkel and Christopher Manning. 2009. Nested named entity recognition. In *EMNLP*, pages 141–150, Singapore.

Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. 2013. Speech recognition with deep recurrent neural networks. In *ICASSP*, pages 6645–6649.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. MIMIC-III, A freely accessible critical care database. *Sci. Data*, 3:160035.

Sarvnaz Karimi, Alejandro Metke-Jimenez, Madonna Kemp, and Chen Wang. 2015a. CADEC: A corpus of adverse drug event annotations. *J Biomed Inform*, 55:73–81.

Sarvnaz Karimi, Chen Wang, Alejandro Metke-Jimenez, Raj Gaire, and Cécile Paris. 2015b. Text and data mining techniques in adverse drug reaction detection. *ACM Comput. Surv.*, 47(4):56:1–56:39.

Arzoo Katiyar and Claire Cardie. 2018. Nested named entity recognition revisited. In *NAACL*, pages 861–871, New Orleans, Louisiana.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *NAACL*, pages 260–270, San Diego, California.

Robert Leaman, Ritu Khare, and Zhiyong Lu. 2015. Challenges in clinical natural language processing for automated disorder normalization. *J Biomed Inform*, 57:28–37.

Robert Leaman, Laura Wojtulewicz, Ryan Sullivan, Annie Skariah, Jian Yang, and Graciela Gonzalez. 2010. Towards internet-age pharmacovigilance: Extracting adverse drug reactions from user posts in health-related social networks. In *BioNLP*, pages 117–125, Uppsala, Sweden.

Yinxia Lou, Yue Zhang, Tao Qian, Fei Li, Shufeng Xiong, and Donghong Ji. 2017. A transition-based joint model for disease named entity recognition and normalization. *Bioinformatics*, 33(15):2363–2371.

Wei Lu and Dan Roth. 2015. Joint mention extraction and classification with mention hypergraphs. In *EMNLP*, pages 857–867, Lisbon, Portugal.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *EMNLP*, pages 1412–1421, Lisbon, Portugal.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *ACL*, pages 1064–1074, Berlin, Germany.

Ryan McDonald, Koby Crammer, and Fernando Pereira. 2005. Flexible text segmentation with structured multilabel classification. In *EMNLP*, pages 987–994, Vancouver, British Columbia, Canada.

Alejandro Metke-Jimenez and Sarvnaz Karimi. 2016. Concept identification and normalisation for adverse drug event discovery in medical forums. In *BM-DID@ISWC*, Kobe, Japan.

Danielle L Mowery, Sumithra Velupillai, Brett R South, Lee Christensen, David Martinez, Liadh Kelly, Lorraine Goeuriot, Noemie Elhadad, Sameer Pradhan, Guergana Savova, and Wendy W Chapman. 2014. Task 2: ShARe/CLEF ehealth evaluation lab 2014. In *CLEF*.

Aldrian Obaja Muis and Wei Lu. 2016. Learning to recognize discontiguous entities. In *EMNLP*, pages 75–84, Austin, Texas.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *NAACL*, pages 2227–2237, New Orleans, Louisiana.

Matthew E. Peters, Sebastian Ruder, and Noah A. Smith. 2019. To tune or not to tune? adapting pretrained representations to diverse tasks. In *RepL4NLP@ACL*, pages 7–14, Florence, Italy.

Sameer Pradhan, Noemie Elhadad, Brett R South, David Martinez, Lee M Christensen, Amy Vogel, Hanna Suominen, Wendy W Chapman, and Guergana K Savova. 2013. Task 1: ShARe/CLEF ehealth evaluation lab 2013. In *CLEF*.

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 *ACL*, pages 5176–5181, Florence, Italy.

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. *J Biomed Inform*, 54:202–212.

Wei Shen, Jianyong Wang, and Jiawei Han. 2015. Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Trans. Knowl. Data Eng.*, 27(2):443–460.

Mohammad Golam Sohrab and Makoto Miwa. 2018. Deep exhaustive model for nested named entity recognition. In *EMNLP*, pages 2843–2849, Brussels, Belgium.

Gabriel Stanovsky, Daniel Gruhl, and Pablo Mendes. 2017. Recognizing mentions of adverse drug reaction in social media using knowledge-infused recurrent models. In *EACL*, pages 142–151, Valencia, Spain.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. brat: a web-based tool for NLP-assisted text annotation. In *EACL*, pages 102–107, Avignon, France.

Buzhou Tang, Hongxin Cao, Yonghui Wu, Min Jiang, and Hua Xu. 2013a. Recognizing clinical entities in hospital discharge summaries using structural support vector machines with word representation features. In *BMC Med Inform Decis Mak*, volume 13, page S1.

Buzhou Tang, Jianglu Hu, Xiaolong Wang, and Qingcai Chen. 2018. Recognizing continuous and discontinuous adverse drug reaction mentions from social media using LSTM-CRF. *Wirel Commun Mob Com*, 2018.

Buzhou Tang, Yonghui Wu, Min Jiang, and Joshua C. Denny. 2013b. Recognizing and encoding disorder concepts in clinical text using machine learning and vector space model. In *CLEF*.

Bailin Wang and Wei Lu. 2019. Combining spans into entities: A neural two-stage approach for recognizing discontiguous entities. In *EMNLP*, pages 6215–6223, Hong Kong, China.

Bailin Wang, Wei Lu, Yu Wang, and Hongxia Jin. 2018a. A neural transition-based model for nested mention recognition. In *EMNLP*, pages 1011–1017, Brussels, Belgium.

Yanshan Wang, Liwei Wang, Majid Rastegar-Mojarad, Sungrim Moon, Feichen Shen, Naveed Afzal, Sijia Liu, Yuqun Zeng, Saeed Mehrabi, Sunghwan Sohn, and Hongfang Liu. 2018b. Clinical information extraction applications: A literature review. *J Biomed Inform*, 77:34–49.

Taro Watanabe and Eiichiro Sumita. 2015. Transition-based neural constituent parsing. In *ACL-IJCNLP*, pages 1169–1179, Beijing, China.

Mingbin Xu, Hui Jiang, and Sedtawut Watcharawittayakul. 2017. A local detection approach for named entity recognition and mention detection. In *ACL*, pages 1237–1247, Vancouver, Canada.