

# Combining Spans into Entities: A Neural Two-Stage Approach for Recognizing Discontiguous Entities

**Bailin Wang**

ILCC, School of Informatics  
University of Edinburgh  
bailin.wang@ed.ac.uk

**Wei Lu**

StatNLP Research Group  
Singapore University of Technology and Design  
luwei@sutd.edu.sg

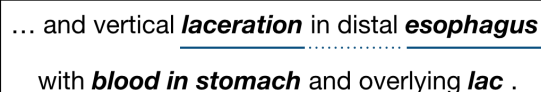
## Abstract

In medical documents, it is possible that an entity of interest not only contains a discontiguous sequence of words but also overlaps with another entity. Entities of such structures are intrinsically hard to recognize due to the large space of possible entity combinations. In this work, we propose a neural two-stage approach to recognize discontiguous and overlapping entities by decomposing this problem into two subtasks: 1) it first detects all the overlapping spans that either form entities on their own or present as segments of discontiguous entities, based on the representation of segmental hypergraph, 2) next it learns to combine these segments into discontiguous entities with a classifier, which filters out other incorrect combinations of segments. Two neural components are designed for these subtasks respectively and they are learned jointly using a shared encoder for text. Our model achieves the state-of-the-art performance in a standard dataset, even in the absence of external features that previous methods used.

## 1 Introduction

Named entity recognition (NER) aims at identifying shallow semantic elements in text and has been a crucial step towards natural language understanding (Tjong Kim Sang and De Meulder, 2003). Extracted entities can facilitate various downstream tasks like question answering (Abney et al., 2000), relation extraction (Mintz et al., 2009; Liu et al., 2017), event extraction (Riedel and McCallum, 2011; Lu and Roth, 2012; Li et al., 2013), and coreference resolution (Soon et al., 2001; Ng and Cardie, 2002; Chang et al., 2013).

The underlying assumptions behind most NER systems are that an entity should contain a contiguous sequence of words and should not overlap with each other. However, such assumptions do



... and vertical *laceration* in distal *esophagus*  
with *blood in stomach* and overlying *lac* .

Figure 1: Entities are highlighted with colored underlines. “*laceration ... esophagus*” and “*stomach ... lac*” contain discontiguous sequence of words and the latter also overlaps with another entity “*blood in stomach*”.

not always hold in practice. First, entities or mentions<sup>1</sup> with overlapping structures frequently exist in news (Doddington et al., 2004) and biomedical documents (Kim et al., 2003). Second, entities can be discontiguous, especially in clinical texts (Pradhan et al., 2014b). For example, Figure 3 shows three entities where two of them are discontiguous (“*laceration ... esophagus*” and “*stomach ... lac*”), and the second discontiguous entity also overlaps with another entity (“*blood in stomach*”).

Such discontiguous entities are intrinsically hard to recognize considering the large search space of possible combinations of entities that have discontiguous and overlapping structures. Muis and Lu (2016a) proposed a hypergraph-based representation to compactly encode discontiguous entities. However, this representation suffers from the ambiguity issue during decoding – one particular hypergraph corresponds to multiple interpretations of entity combinations. As a result, it resorted to heuristics to deal with such an issue.

Motivated by their work, we take a novel approach to resolve the ambiguity issue in this work. Our core observation is that though it is hard to exactly encode the exponential space of all possible discontiguous entities, recent work on extracting overlapping structures (Wang and Lu, 2018) can be employed to efficiently explore the space

<sup>1</sup>Mentions are defined as references to entities that could be named, nominal or pronominal (Florian et al., 2004).

of all the span combinations of discontinuous entities. Based on this observation, we decompose the problem of recognizing discontinuous entities into two subtasks: 1) *segment extraction*: learning to detect all (potentially overlapping) spans that either form entities on their own or present as parts of a discontinuous entity; 2) *segment merging*: learning to form entities by merging certain spans into discontinuous entities.

Our contributions are summarized as follows:

- By decomposing the problem of extracting discontinuous entities into two subtasks, we propose a two-stage approach that does not have the ambiguity issue.
- Under this decomposition, we design two neural components for these two subtasks respectively. We further show that the joint learning setting where the two components use a shared text encoder is beneficial.
- Empirical results show that our system achieves a significant improvement compared with previous methods, even in the absence of external features that previous methods used.<sup>2</sup>

Though we only focus on discontinuous entity recognition in this work, our model may find applications in other tasks that involve discontinuous structures, such as detecting gappy multiword expressions (Schneider et al., 2014).

## 2 Related Work

The task of extracting overlapping entities has long been studied (Zhang et al., 2004; Zhou et al., 2004; Zhou, 2006; McDonald et al., 2005; Alex et al., 2007; Finkel and Manning, 2009; Lu and Roth, 2015; Muis and Lu, 2017). As neural models (Collobert et al., 2011; Lample et al., 2016; Huang et al., 2015; Chiu and Nichols, 2016; Ma and Hovy, 2016) are proven effective for NER, there have been several neural systems recently proposed to handle entities of overlapping structures (Ju et al., 2018; Katiyar and Cardie, 2018; Wang et al., 2018; Sohrab and Miwa, 2018; Wang and Lu, 2018; Straková et al., 2019; Lin et al., 2019; Fisher and Vlachos, 2019). Our system is based on the model of neural segmental hypergraphs (Wang and Lu, 2018) which encodes all the

possible combinations of overlapping entities using a compact hypergraph representation without ambiguity. Note that other system for extracting overlapping structures can also fit into our two-stage system.

For discontinuous and overlapping entity recognition, Tang et al. (2013); Zhang et al. (2014); Xu et al. (2015) extended the BIO tagging scheme to encode such complex structures so that traditional linear-chain CRF (Lafferty et al., 2001) can be employed. However, the model suffers greatly from ambiguity during decoding due to the use of the extended tagset. Muis and Lu (2016a) proposed a hypergraph-based representation to reduce the level of ambiguity. Essentially, these systems trade expressiveness for efficiency: they inexactly encoded the whole space of discontinuous entities with ambiguity for training, and then relied on some heuristics to handle the ambiguity during decoding.<sup>3</sup> Considering it is intrinsically hard to exactly identify discontinuous entities in one stage using a structured model, our work tries to decompose the task into two sub-tasks to resolve the ambiguity issue.

This task is also related to joint entity and relation extraction (Kate and Mooney, 2010; Li and Ji, 2014; Miwa and Sasaki, 2014) where the discontinuous entities can be viewed as relation links between segments. The major difference is that discontinuous entities require explicitly modeling overlapping entities and linking multiple segments.

## 3 Model

Our goal is to extract a set of entities that may have overlapping and discontinuous structures given a natural language sentence. We use  $x = x_1 \dots x_{|x|}$  to denote a sentence, and use  $y = \{[b_{i:j} \dots b_{m:n}]^k\}$  to denote a set of discontinuous entities where each entity of type  $k$  contains a list of spans, e.g.,  $b_{i:j}$  and  $b_{m:n}$ , with subscripts indicating the starting and ending positions of the span. Hence, this task can be viewed as extracting and labelling a sequence of spans as an entity.

Our two-stage approach first extracts spans of interest like  $b_{i:j}$ , which are parts of discontinuous entities. Then it merges these extracted spans into discontinuous entities. In the more general setting where discontinuous entities are typed, our

<sup>2</sup>Our code is available at [https://github.com/berlino/disco\\_em19](https://github.com/berlino/disco_em19).

<sup>3</sup>We will briefly introduce them and their heuristics later as our baselines in our experiments for comparison.

approach is designed to jointly extract and label the spans at the first stage, then only merge the spans of the same type at the second stage. We call the intermediate typed span  $b_{i:j}^k$  a *segment* in the rest of the paper.

Formally, our model aims at maximizing the conditional probability  $p(\mathbf{y}|\mathbf{x})$ , which is decomposed as:

$$p(\mathbf{y}|\mathbf{x}) = p(\mathbf{s}|\mathbf{x})p(\mathbf{y}|\mathbf{s}, \mathbf{x}) \quad (1)$$

where  $\mathbf{s} = \{b_{i:j}^k\}$  denotes the set of segments that leads to  $\mathbf{y}$  through a specific combination.<sup>4</sup> That is, we divide the problem of extracting discontinuous entities into two subtasks, namely *segment extraction* and *segment merging*.

### 3.1 Segment Extraction

The entity segments  $\mathbf{s}$  of interest in a given sentence could also overlap with each other. For example, in Figure 3 the entity “*blood in stomach*” contains another segment “*stomach*”. To make our model capable of extracting such overlapping segment combinations, we employ the model of neural segmental hypergraphs from Wang and Lu (2018), which uses a hypergraph-based representation to encode all the possible combinations of segments without ambiguity. Specifically, the segmental hypergraphs adopt a log-linear approach to model the conditional probability of each segment combination for a given sentence:

$$p(\mathbf{s}|\mathbf{x}) = \frac{\exp f(\mathbf{x}, \mathbf{s})}{\sum_{\mathbf{s}'} \exp f(\mathbf{x}, \mathbf{s}')} \quad (2)$$

where  $f(\mathbf{x}, \mathbf{s})$  is the score function for any pair of input sentence  $\mathbf{x}$  and output segment combination  $\mathbf{s}$ .

In segmental hypergraphs, each segment combination  $\mathbf{s}$  corresponds to a hyperpath. Following Wang and Lu (2018), the score for a hyperpath is the sum of the scores for each hyperedge, which are based on the word-level and span-level representations through LSTM (Graves and Schmidhuber, 2005):

$$\mathbf{h}_i^w = [\text{biLSTM}_1(\mathbf{x}_0, \dots, \mathbf{x}_n)]_i \quad (3)$$

$$\mathbf{h}_{i:j}^s = \text{biLSTM}_2(\mathbf{h}_i^w, \dots, \mathbf{h}_j^w) \quad (4)$$

where  $\mathbf{x}_k$  is the corresponding word embedding for word  $x_k$ ,  $\mathbf{h}_i^w$  denotes the representation for the

<sup>4</sup>We note that each  $\mathbf{y}$  corresponds to one unique  $\mathbf{s}$ .

	# <i>sents</i>	# <i>entities</i> (%)			# <i>o.l.</i> (%)
		1 <i>segment</i>	2 <i>segments</i>	3 <i>segments</i>	
Train	534	544 (46)	607 (51)	44 (4)	205 (17)
Dev	303	357 (45)	421 (53)	18 (2)	240 (30)
Test	430	584 (48)	610 (50)	16 (1)	327 (27)

Table 1: Statistics of the dataset. *o.l.*: overlapping entities, *sents*: sentences.

$i$ -th word and  $\mathbf{h}_{i:j}^s$  denotes the representation for the span from the  $i$ -th to the  $j$ -th word.

On top of the segmental hypergraph representation, the partition function which is the denominator of Equation 2 can be computed using dynamic programming. The inference algorithm has a quadratic time complexity in the number of words, which can be further reduced to linear time complexity if we introduce the maximal length  $c$  of a segment. We regard  $c$  as a hyperparameter.

### 3.2 Segment Merging

Given a set of segments, our next subtask is to merge them into entities. First, we enumerate all the valid segment combinations, denoted as  $\mathbf{E}$ , based on the assumption that the segments in the same entity should have the same type and not overlap with each other. Our model then independently decides whether each valid segment combination forms an entity. We call these valid segment combinations *entity candidates*. For brevity, let us use  $\mathbf{t}^k$  to denote an entity candidate  $[b_{i:j} \dots b_{m:n}]^k$  where each segment like  $b_{i:j}^k$  belongs to  $\mathbf{s}$ .

Formally, given segments  $\mathbf{s}$ , the probability of generating entities  $\mathbf{y}$  can be represented as:

$$p(\mathbf{y}|\mathbf{s}, \mathbf{x}) = \prod_{\mathbf{t}^k \in \mathbf{E}} (\mathbb{1}[\mathbf{t}^k \in \mathbf{y}]p(\mathbf{t}^k \in \mathbf{y}) + \mathbb{1}[\mathbf{t}^k \notin \mathbf{y}](1 - p(\mathbf{t}^k \in \mathbf{y}))) \quad (5)$$

where  $\mathbb{1}$  is an indicator function. We use a binary classifier to model  $p(\mathbf{t}^k \in \mathbf{y})$ .

To capture the interactions between segments within the same combination, we employ yet another LSTM on top of segments as follows:

$$\mathbf{h}_{\mathbf{t}^k}^e = \text{biLSTM}_3(\mathbf{h}_{i:j}^s, \dots, \mathbf{h}_{m:n}^s) \quad (6)$$

where  $\mathbf{h}_{\mathbf{t}^k}^e$  denotes the representation of the segment combination  $\mathbf{t}^k$ , which then serves as a feature vector for a binary classifier to determine whether it is an entity. Note that we reuse the span

representation from Equation 4, meaning that encoder for words and spans are shared in both segment extraction and merging.

The binary classifier for each  $t^k$  in Equation 5 is computed as:

$$p(t^k \in y) = \text{sigmoid}(\mathbf{W} \cdot \text{relu}(\mathbf{h}_{t^k}^e) + b) \quad (7)$$

where we use a rectified linear unit (ReLU) (Glorot et al., 2011) and a linear layer, parameterized by  $\mathbf{W}$  and  $b$ , to map the representation from Equation 6 to a scalar score. This score is normalized into a distribution by the sigmoid function.

In the joint model, we stack three separate LSTMs to encode text at different levels from words to spans, then discontinuous entities. Intuitively, the word and span level LSTM try to capture the lower-level information for segment extraction while the entity level LSTM captures the higher-level information for segment merging.

### 3.3 Learning and Decoding

For a dataset  $\mathcal{D}$  consisting of sentence-entities pairs  $(x, y)$ , our objective is to minimize the negative log-likelihood as follows:

$$\mathcal{L}(\theta) = - \sum_{(x_i, y_i) \in \mathcal{D}} \log p(y_i | x_i) + \frac{\lambda}{2} \|\theta\|^2 \quad (8)$$

where  $\theta$  denotes model parameters and  $\lambda$  is the  $\ell_2$  coefficient.  $p(y_i | x_i)$  is computed by  $p(s_i | x_i) p(y_i | s_i, x_i)$  where  $s_i$  is inferred from  $y_i$ .

During the decoding stage, the system first predicts the most probable segments from the neural segmental hypergraph by  $\hat{s} = \arg \max_{s'} p(s' | x)$ . Then it feeds the prediction to the next stage of merging segments and outputs the discontinuous entities by  $\hat{y} = \arg \max_{y'} p(y' | \hat{s}, x)$ .

## 4 Experiments

### 4.1 Setup

**Data** We evaluated our model on the task of recognizing mentions in clinical text from ShARe/CLEF eHealth Evaluation Lab (SHEL) 2013 (Suominen et al., 2013) and SemEval-2014 (Pradhan et al., 2014a). The task is defined to extract mentions of *disorders* from clinical documents according to the Unified Medical Language System (UMLS). The original dataset only has a small percentage of discontinuous entities, making it not suitable for comparing the effectiveness of different models when handling discontinuous

entities. Following Muis and Lu (2016a), we use a subset of the original data where each sentence contains at least one discontinuous entity.

We split the dataset according to the setting of SemEval 2014. Statistics are shown in Table 1. In this subset, 53.6% of entities are discontinuous. Overlapping entities also frequently appear. Since an entity has three segments at most, we make the constraint that an entity candidate has no more than three segments during segment merging.

Note that all entities hold the same type of *disorder* in this dataset. Our model is intrinsically able to handle discontinuous entities of multiple types. Recall that segments are typed as  $b_{i:j}^k$  during segment extraction, and only segments of the same type can be merged into an entity  $t^k$  where  $k$  indicates the entity type. To assess its ability to deal with multiple entity types, we conducted a further analysis (see section 4.2).

**Hyperparameters** We use the pretrained word embeddings from Chiu et al. (2016) which are trained on the PubMed corpus. A dropout layer (Srivastava et al., 2014) is used after each word is mapped to its embedding. The dropout rate and the number of hidden units in LSTMs are tuned based on the performance on the development set. We set the maximal length of a segment to be 6 during segment extraction. Our model is trained with Adam (Kingma and Ba, 2014).<sup>5</sup>

**Baselines** The first baseline we consider is to extend the traditional BIO tagging scheme to seven tags following Tang et al. (2013). With this tagging scheme, each word in a sentence is assigned a label. Then a linear-chain CRF is built to model the sequence labelling process. The next baseline is a hypergraph-based method by Muis and Lu (2016a). It encodes each entity combination into a directed graph based on six types of nodes; each has its specific semantics.

Since these baselines are both ambiguous, heuristics are required during decoding. Following Muis and Lu (2016a), we explored two heuristics: given a model’s ambiguous output, either a tag sequence or a hypergraph, the “enough” heuristic finds the minimal set of entities that corresponds to it, while “all” decodes the union of all the possible set of entities. Please refer to Muis and Lu (2016a) for details. We also describe them in the Appendix for self-containedness.

<sup>5</sup>See Appendix C for the full hyperparameters.



	Model	$P$	$R$	$F_1$
Non-Neural	CRF (enough)	54.7	41.2	47.0
	CRF (all)	15.2	44.9	22.7
	Graph (enough)	<b>76.9</b>	40.1	52.7
	Graph (all)	76.0	40.5	52.8
	<i>Our model</i>	76.3	41.4	53.7
Neural	CRF (enough)	43.7	54.3	48.4
	CRF (all)	15.7	55.8	24.5
	<i>Our model</i>	48.4	<b>66.5</b>	<b>56.1</b>
	<i>w.o. shared encoder</i>	46.2	65.1	54.0

Table 2: Main results. Graph: the hypergraph based model by Muis and Lu (2016a). “enough” and “all” denotes the heuristics used in ambiguous model. *w.o. shared encoder*: without using shared encoder.

We compare our approach to these baselines in two settings. In the non-neural setting, we compare models using the same set of handcrafted features, including external features from POS tagger and Brown cluster following (Muis and Lu, 2016a). In the neural setting, we implement a linear-chain CRF model using the same neural encoder. We are trying to see our model can perform better in both settings. Note that all neural models in our experiments do not leverage any handcrafted features.

## 4.2 Results and Analysis

The main results are listed in Table 2. In both non-neural and neural settings, our model achieves the better result in terms of  $F_1$  compared with other baselines, revealing the effectiveness of our methodology of decomposing the task into two stages. Our neural model achieves the best performance even without using any external handcrafted features.

We also assess the performance when our model uses separate encoders for segment extraction and merging. From the results, we observe that the setting of using a shared encoder is very beneficial for our two-stage system.

Compared with non-neural models, neural models are better in terms of  $F_1$ , both for CRF and our models. The gain mostly comes from the ability to recall more entities. Handcrafted features in non-neural models lead to high precisions but do not seem to be general enough to recall most entities.

The “enough” heuristic works better than “all” in most cases. Hence we use it for evaluating models’ ability in handling multiple entity types.

**Handling Multiple Entity Types** To assess the effectiveness of handling entities of multi-

	Model	$P$	$R$	$F_1$
Non-Neural	CRF (enough)	55.3	37.4	44.6
	Graph (enough)	<b>67.3</b>	37.5	48.2
Neural	CRF (enough)	41.6	52.3	46.3
	<i>Our model</i>	43.3	<b>65.8</b>	<b>52.2</b>

Table 3: Results on handling multiple entity types.

ple types, we further categorize each entity into three types based on its Concept Unique Identifier (CUI), following Muis and Lu (2016a).<sup>6</sup> In this setting, segments are jointly extracted and labelled using these three categories during segment extraction. During segment merging, an entity candidate can only contain segments of the same type during merging.

The results are listed in Table 3. Our neural model again achieves the best performance among all models in terms of  $F_1$ . Compared with neural CRF, our model is significantly better at recalling entities. Similar to the previous observation, the neural encoder consistently boosts the performance of the CRF by recalling more entities, compared with its non-neural counterpart.

## 5 Conclusion and Future Work

In this work, we propose a neural two-stage approach for recognizing discontinuous entities, which learns to extract and merge segments jointly without suffering from ambiguity issue. Empirically, it achieves a significant improvement compared with previous methods that rely heavily on handcrafted features.

During training, the classifier of merging segments is only exposed to correct segments, making it unable to recover from errors of segment extraction during decoding. This issue is similar to *exposure bias* (Wiseman and Rush, 2016) and it might be beneficial if the classifier of segment merging is exposed to incorrect segments during training. We leave this for future work.

## Acknowledgements

We would like to thank the anonymous reviewers for their valuable comments. Wei Lu is supported by Singapore Ministry of Education Academic Research Fund (AcRF) Tier 2 Project MOE2017-T2-1-156, and is partially supported by SUTD project PIE-SGP-AI-2018-01.

<sup>6</sup>Note that the label of CUI is not available for all entities. Entities without CUI are assigned with a default *NULL* label.

## References

- Steven Abney, Michael Collins, and Amit Singhal. 2000. Answer extraction. In *Proc. of the sixth conference on applied natural language processing*.
- Beatrice Alex, Barry Haddow, and Claire Grover. 2007. Recognising nested named entities in biomedical text. In *Proc. of BioNLP*.
- Kai-Wei Chang, Rajhans Samdani, and Dan Roth. 2013. A constrained latent variable model for coreference resolution. In *Proc. of EMNLP*.
- Billy Chiu, Gamal Crichton, Anna Korhonen, and Sampo Pyysalo. 2016. How to train good word embeddings for biomedical nlp. In *Proc. of the 15th Workshop on Biomedical Natural Language Processing*.
- Jason PC Chiu and Eric Nichols. 2016. Named entity recognition with bidirectional lstm-cnns. *TACL*, 4:357–370.
- Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *JMLR*, 12(Aug):2493–2537.
- George R Doddington, Alexis Mitchell, Mark A Przybicki, Lance A Ramshaw, Stephanie Strassel, and Ralph M Weischedel. 2004. The automatic content extraction (ace) program-tasks, data, and evaluation. In *Proc. of LREC*.
- Jenny Rose Finkel and Christopher D Manning. 2009. Nested named entity recognition. In *Proc. of EMNLP*.
- Joseph Fisher and Andreas Vlachos. 2019. Merge and label: A novel neural network architecture for nested NER. In *Proc. of ACL*, pages 5840–5850.
- R. Florian, H. Hassan, A. Ittycheriah, H. Jing, N. Kambhatla, X. Luo, N. Nicolov, and S. Roukos. 2004. A statistical model for multilingual entity detection and tracking. In *Proc. of HLT-NAACL*.
- Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Deep sparse rectifier neural networks. In *Proc. of AISTATS*.
- Alex Graves and Jürgen Schmidhuber. 2005. Frame-wise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, 18(5-6):602–610.
- Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.
- Meizhi Ju, Makoto Miwa, and Sophia Ananiadou. 2018. A neural layered model for nested named entity recognition. In *Proc. of NAACL-HLT*.
- Rohit J Kate and Raymond J Mooney. 2010. Joint entity and relation extraction using card-pyramid parsing. In *Proc. of CoNLL*.
- Arzoo Katiyar and Claire Cardie. 2018. Nested named entity recognition revisited. In *Proc. of NAACL-HLT*.
- J-D Kim, Tomoko Ohta, Yuka Tateisi, and Jun’ichi Tsujii. 2003. Genia corpus—a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(suppl\_1):i180–i182.
- Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. In *Proc. of ICLR*.
- John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. of ICML*.
- Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proc. of NAACL-HLT*.
- Qi Li and Heng Ji. 2014. Incremental joint extraction of entity mentions and relations. In *Proc. of ACL*.
- Qi Li, Heng Ji, and Liang Huang. 2013. Joint event extraction via structured prediction with global features. In *Proc. of ACL*.
- Hongyu Lin, Yaojie Lu, Xianpei Han, and Le Sun. 2019. Sequence-to-nuggets: Nested entity mention detection via anchor-region networks. In *Proc. of ACL*, pages 5182–5192.
- Liyuan Liu, Xiang Ren, Qi Zhu, Shi Zhi, Huan Gui, Heng Ji, and Jiawei Han. 2017. Heterogeneous supervision for relation extraction: A representation learning approach. In *Proc. of EMNLP*.
- Wei Lu and Dan Roth. 2012. Automatic event extraction with structured preference modeling. In *Proc. of ACL*.
- Wei Lu and Dan Roth. 2015. Joint mention extraction and classification with mention hypergraphs. In *Proc. of EMNLP*.
- Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. In *Proc. of ACL*.
- Ryan McDonald, Koby Crammer, and Fernando Pereira. 2005. Flexible text segmentation with structured multilabel classification. In *Proc. of HLT-EMNLP*.
- Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proc. of ACL-IJCNLP*.
- Makoto Miwa and Yutaka Sasaki. 2014. Modeling joint entity and relation extraction with table representation. In *Proc. of EMNLP*.

- Aldrian Obaja Muis and Wei Lu. 2016a. Learning to recognize discontinuous entities. In *Proc. of EMNLP*.
- Aldrian Obaja Muis and Wei Lu. 2016b. Supplementary material for recognizing overlapping mentions with mention separators.
- Aldrian Obaja Muis and Wei Lu. 2017. Labeling gaps between words: Recognizing overlapping mentions with mention separators. In *Proc. of EMNLP*.
- Vincent Ng and Claire Cardie. 2002. Improving machine learning approaches to coreference resolution. In *Proc. of ACL*.
- Sameer Pradhan, Noémie Elhadad, Wendy Chapman, Suresh Manandhar, and Guergana Savova. 2014a. Semeval-2014 task 7: Analysis of clinical text. In *Proc. of SemEval*.
- Sameer Pradhan, Noémie Elhadad, Brett R South, David Martinez, Lee Christensen, Amy Vogel, Hanna Suominen, Wendy W Chapman, and Guergana Savova. 2014b. Evaluating the state of the art in disorder recognition and normalization of the clinical narrative. *Journal of the American Medical Informatics Association*, 22(1):143–154.
- S. Riedel and A. McCallum. 2011. Fast and robust joint models for biomedical event extraction. In *Proc. of EMNLP*.
- Nathan Schneider, Emily Danchik, Chris Dyer, and Noah A Smith. 2014. Discriminative lexical semantic segmentation with gaps: running the mwe gamut. *TACL*, 2:193–206.
- Mohammad Golam Sohrab and Makoto Miwa. 2018. Deep exhaustive model for nested named entity recognition. In *Proc. of EMNLP*.
- Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational linguistics*, 27(4):521–544.
- Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *JMLR*, 15(1):1929–1958.
- Jana Straková, Milan Straka, and Jan Hajic. 2019. Neural architectures for nested NER through linearization. In *Proc. of ACL*, pages 5326–5331.
- Hanna Suominen, Sanna Salanterä, Sumithra Velupillai, Wendy W Chapman, Guergana Savova, Noémie Elhadad, Sameer Pradhan, Brett R South, Danielle L Mowery, Gareth JF Jones, et al. 2013. Overview of the share/clef ehealth evaluation lab 2013. In *Proc. of International Conference of the Cross-Language Evaluation Forum for European Languages*.
- Buzhou Tang, Hongxin Cao, Yonghui Wu, Min Jiang, and Hua Xu. 2013. Recognizing clinical entities in hospital discharge summaries using structural support vector machines with word representation features. In *Proc. of BMC medical informatics and decision making*.
- Erik F Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proc. of CoNLL*.
- Bailin Wang and Wei Lu. 2018. Neural segmental hypergraphs for overlapping mention recognition. In *Proc. of EMNLP*.
- Bailin Wang, Wei Lu, Yu Wang, and Hongxia Jin. 2018. A neural transition-based model for nested mention recognition. In *Proc. of EMNLP*.
- Sam Wiseman and Alexander M. Rush. 2016. Sequence-to-sequence learning as beam-search optimization. In *Proc. of EMNLP*.
- Jun Xu, Yaoyun Zhang, Jingqi Wang, Yonghui Wu, Min Jiang, Ergin Soysal, and Hua Xu. 2015. Uth-ccb: the participation of the semeval 2015 challenge-task 14. In *Proc. of SemEval*.
- Jie Zhang, Dan Shen, Guodong Zhou, Jian Su, and Chew-Lim Tan. 2004. Enhancing hmm-based biomedical named entity recognition by studying special phenomena. *Journal of biomedical informatics*, 37(6):411–422.
- 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 *Proc. of SemEval*.
- Guodong Zhou. 2006. Recognizing names in biomedical texts using mutual information independence model and svm plus sigmoid. *International Journal of Medical Informatics*, 75(6):456–467.
- Guodong Zhou, Jie Zhang, Jian Su, Dan Shen, and Chewlim Tan. 2004. Recognizing names in biomedical texts: a machine learning approach. *Bioinformatics*, 20(7):1178–1190.

## A Segment Extraction

Neural segmental hypergraphs (Wang and Lu, 2018) were proposed for modeling overlapping structures in entity mentions. We directly adopt their approach to model the segments of overlapping structures. Note that our segment also holds the information of entity type, so the resulting system for segment extraction can also be viewed as performing sub-mention recognition. Next, we illustrate how the segmental hypergraph encodes the overlapping segments by a concrete example.

For brevity, we only show the example that is annotated with one entity type, and it is able to be trivially extended to the case of multiple entity types.

Given a phrase “He had blood in his mouth and on his tongue”, there exist two disorder mentions: ‘*blood in his mouth*’ and ‘*blood ... on his tongue*’ where the second mention has a discontinuous sequence of words. Our two-stage approach first extracts segments that lead to these two mentions. In this example, the segments consist of ‘*blood*’, ‘*blood in his mouth*’ and ‘*on his tongue*’. We observe that the first two segments overlap with each other.

Segmental hypergraph encodes this segment combination based on five types of nodes:

- $\mathbf{A}_i$  encodes all segments that start with the  $i$ -th or a later word
- $\mathbf{E}_i$  encodes all segments that start exactly with the  $i$ -th word
- $\mathbf{T}_i^k$  represents all segments of type  $k$  starting with the  $i$ -th word
- $\mathbf{I}_{i,j}^k$  represents all segments of type  $k$  that contain the  $j$ -th word and start with the  $i$ -th word
- $\mathbf{X}$  marks the end of a segment.

Each segment can be expressed in terms of these five nodes and corresponds with a path in the segmental hypergraph. As a result, each segment combination corresponds with a *hyperpath* where hyperedges are designated to connect multiple nodes so as to model overlapping segments. Figure 2 shows such a hyperpath for the segment combination in our example phrase. Since we only have one entity type in this example, we eliminate the superscript  $k$  in  $\mathbf{T}$  and  $\mathbf{I}$  nodes that indicates the information of entity type.

Starting from the third word ‘blood’, there exist two segments ‘*blood*’ and ‘*blood in his mouth*’. The brown hyperedge with the parent node being  $\mathbf{I}_{3,3}$  is responsible for connecting these two overlapping segments. This hyperedge means that there exists a segment that ends at the third word (the link from  $\mathbf{I}_{3,3}$  to  $\mathbf{X}$ ) and there also exists a segment that continues to the next word (the link from  $\mathbf{I}_{3,3}$  to  $\mathbf{I}_{3,4}$ ). The segment ‘*on his tongue*’ is directly mapped to the path from  $\mathbf{T}_8$  to  $\mathbf{X}$ .

The score for each hyperpath is the sum of the scores that are computed over each hyperedge.

Since  $\mathbf{T}$  nodes encode word-level information and  $\mathbf{I}$  nodes encode span-level information, two LSTMs are employed to capture the interactions at both word-level and span-level respectively. We use their original implementation that is publicly available<sup>7</sup>.

## B Heuristics for Handling Ambiguity

This section tries to explain the two heuristics “enough” and “all” when ambiguous tag sequences occur. We use the extended BIO tagging scheme (Tang et al., 2013; Muis and Lu, 2016a) for example.

To encode the three discontinuous entities in Figure 3, this tagset has seven tags:

- **B/I:** Beginning and Inside of contiguous entities
- **BH/IH:** Beginning and Inside of *head* where head refers to segments shared by multiple discontinuous entities.
- **BD/ID:** Beginning and Inside of *body* where body refers to segments that are not shared across entities.
- **O:** Outside of entities.

The resulting tag sequence is shown in Figure 4. Since this tagging scheme cannot model the correspondence between different tags, tagging sequences are very likely to have multiple interpretations. For instance, it is not clear that “laceration” should be combined with “esophagus” or with “stomach”.

The “all” heuristics extracts all the possible entities that could exist in the tagging sequence. In this case, “all” heuristics will produce “*laceration ... esophagus*”, “*stomach ... lac*”, “*blood in stomach*”, “*laceration ... lac*”, “*esophagu ... lac*”, “*laceration ... esophagus ... lac*”, “*laceration ... stomach*”, “*esophagus ... stomach*”, “*laceration ... stomach ... lac*”, “*esophagus ... stomach...lac*”.

The “enough” heuristics tries to find the minimal set of entities that corresponds to this tagging sequence. In this case, “enough” heuristics would produce at least three entities like: “*laceration ... esophagus*”, “*stomach ... lac*” and “*blood in stomach*”; “*laceration ... lac*”, “*blood in stomach*” and

<sup>7</sup><https://github.com/berlino/overlapping-ner-em18>



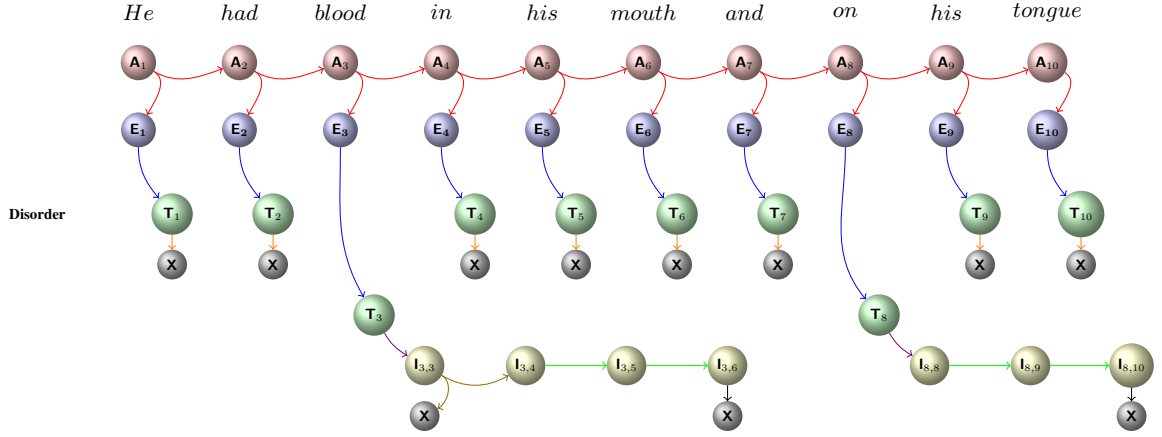


Figure 2: A hyperpath that encodes three mention segments: ‘blood’, ‘blood in his mouth’ and ‘on his tongue’.

... and vertical laceration in distal esophagus  
 with blood in stomach and overlying lac .

Figure 3: Entities are highlighted with colored underlines. “laceration ... esophagus” and “stomach ... lac” contain discontinuous sequence of words and the latter also overlaps with another entity “blood in stomach”.

... and vertical laceration in distal esophagus  
 with blood in stomach and overlying lac .

Figure 4: Entities annotated using seven tags.

“esophagus ... stomach”. We make further constraints to generate only one combination following Muis and Lu (2016b).

## C Hyperparameters

The hyperparameters used in our neural two-stage model are listed in Table 4. Since the size of our

dataset is relatively small, the dropout is crucial to prevent overfitting considering that the pre-trained word embeddings have the dimension of 200. The length of most segments is not greater than 6, so we set the maximal length  $c$  to be 6 to improve the efficiency of segment extraction.

We also tried to incorporate a character-level component (Lample et al., 2016) to capture morphological and orthographic information. However, it does not have a significant effect on the performance in term of  $F_1$ .

word embedding dim	200
LSTM(word) hidden size	128
LSTM(span) hidden size	128
LSTM(entity) hidden size	64
maximal length $c$	6
dropout	0.8
$l_2$	0.0001

Table 4: Hyperparameters of our joint model.