Supplementary Paper for Neural Segmental Hypergraphs for Overlapping Mention Recognition

Bailin Wang
University of Massachusetts
Amherst
bailinwang@cs.umass.edu

Wei Lu
Singapore University of Technology
and Design
luwei@sutd.edu.sg

Abstract

This paper presents some details for neural segmental hypergraphs(SH) (Wang and Lu, 2018). The first section focuses on the construction of SH, detailed proof of the structural ambiguity free theorem and the inference algorithms of SH. The second section gives some details of our experiments including the full statistics of our datasets, handcrafted features used in SH(-NN), the choices of hyperparameters and some complete experiment results.

1 Segmental Hypergraph

1.1 Representation

We illustrate how our segmental hypergraph encodes the structures of overlapping mentions by a concrete example.

First, we introduce the construction of the complete segmental hypergraph which is used to encode all the combinations of mentions in a sentence. The intuition is that at each word, the segmental hypergraph compactly represents all segments starting from this word. To jointly model the labels of these segments(mentions), multiple paths are created with each corresponding to a mention type. Hence, there exist three kinds of combinations in the hypergraph. First, the mentions of the same type k and the same left boundary are compactly encoded hyperedges from $\mathbf{I}$ and $\mathbf{X}$ nodes. Second, we combine the mentions of different types by the hyperedge $\{\mathbf{E} \rightarrow (\mathbf{T}^1, \ldots, \mathbf{T}^m)\}$. Finally, we combine the mentions from different positions by the hyperedge between $\mathbf{A}$ and $\mathbf{E}$ nodes.

In Figure 1, we show the complete expansion of the first word and expansions at other positions can be figured out analogously. If we make a restriction on the maximal length, say 6, then the restricted segmental hypergraph shrinks to Figure 2. Under this setting, the longest possible mention starting from the first word should only be "*Israeli UN Ambassador, Yehuda Lancry*".

An instance of mention combinations corresponds to a hyperpath in the segmental hypergraph. Specifically, from the root node, $\mathbf{A}_1$, we should make a decision of choosing which hyperedge to take recursively. Figure 3 gives the partial hyperpath encoding the three mentions from the first word. Using this example, we present how the hyperpath is constructed. First, the mention "*Israeli UN Ambassador*" and "*Israeli UN Ambassador, Yehuda Lancry*" belong to the same mention type PER. They are compactly represented by a path from node $\mathbf{T}_1^2$. This path has two leaf nodes $\mathbf{X}$ with each corresponding with a mention. Analogously, "*Israeli*" (GPE) can be encoded with the path from $\mathbf{T}_1^1$. Note that we don't use a NULL label to indicate a span without any type. Instead, in this example, the hyperedge between $\mathbf{T}_1^3$ and the leaf node $\mathbf{X}$ indicates that there is no ORG mentions starting from the first word. Next, we combine the mentions of different types by the hyperedge $\{\mathbf{E}_1 \rightarrow (\mathbf{T}_1^1, \mathbf{T}_1^2, \mathbf{T}_1^3))\}$. Finally, we combine the mentions from the next position recursively by the hyperedge $\{\mathbf{A}_1 \rightarrow (\mathbf{A}_1, \mathbf{E}_1)\}$.

1.2 Proof of Structural Ambiguity Free

Theorem 1.1. (Structural Ambiguity Free) *For any sentence, there exists a segmental hypergraph $\mathcal{G}_c = (\mathcal{V}_c, \mathcal{E}_c)$ such that any combination of mentions in the sentence corresponds to exactly one hyperpath $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ within the complete segmental hypergraph, and each hyperpath corresponds to exactly one combination of mentions in a sentence.*

Proof We first prove that each mention combination can be encoded with a unique hyperpath. Assume we are given a combination of mentions in

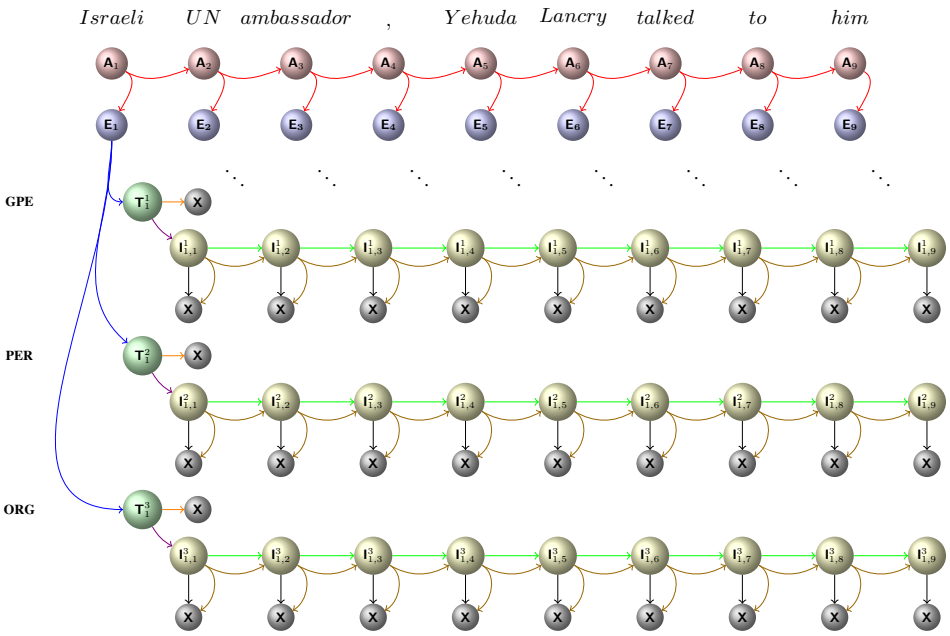

Figure 1: Partial segmental hypergraph that shows the expansions from the first word. We use three mentions types for illustration: GPE(Geo-Political Entity), PER(Person), ORG(Organization).

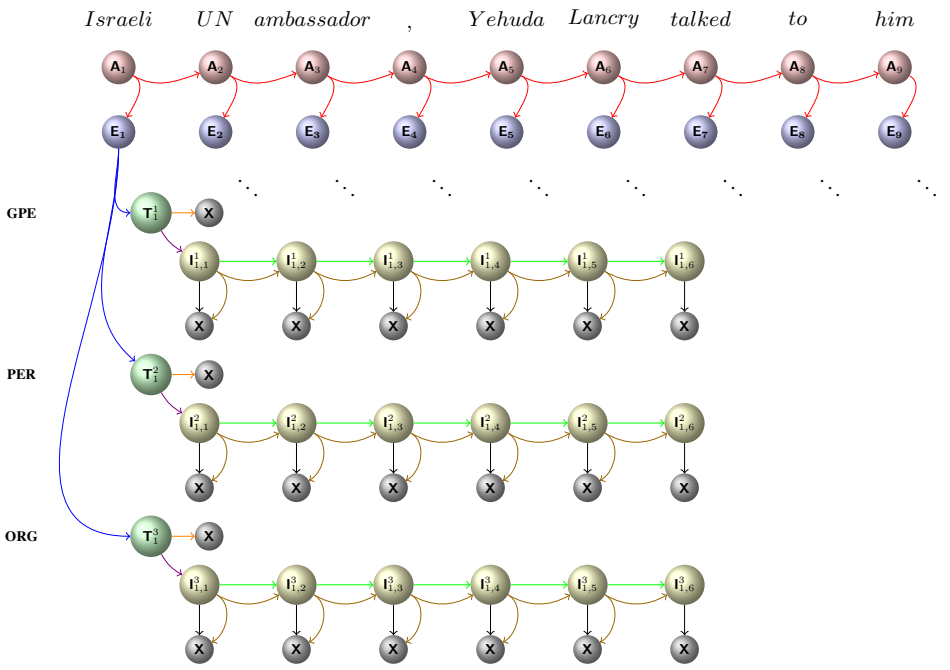

Figure 2: Partial segmental hypergraph with the restriction of maximal mention length ($c = 6$). From the first word, as shown, it can only explore to the sixth word at most.

a sentence, we can construct a hyperpath by visiting nodes as we defined in the paper. We note that the first two types of hyperedges (from **A** and **E** nodes) are always present in any hyperpath. We just need to specify the remaining hyperedges. The hyperedges can be selected based on the position of the mentions. For example, if there exists a mention of type k that starts at a position i, we select the hyperedge that connects $\mathbf{T}_i^k$ and $\mathbf{I}_{i,i}$. Otherwise, we select the hyperedge between $\mathbf{T}_i^k$ and $\mathbf{X}$. Similarly for the other hyperedges. This process forms a unique hyperpath encoding a specific kind of mention combination.

Next, we show why the converse of the above

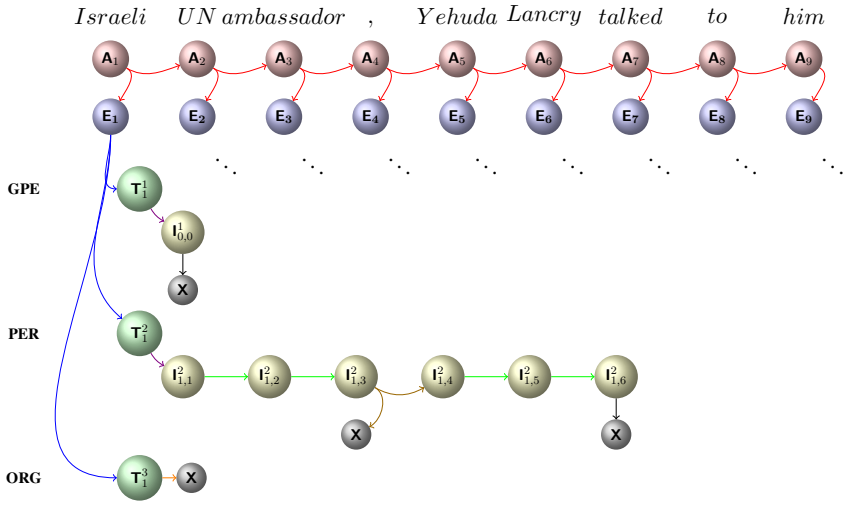

Figure 3: A partial hyperpath for encoding three mentions from the first word: "*Israeli*"(GPE), "*Israeli UN Ambassador*"(PER) and "*Israeli UN Ambassador, Yehuda Lancry*"(PER).

is also true. For a given hyperpath. Consider any hyperedge in the hyperpath that contains both $\mathbf{I}^k_{i,j}$ and $\mathbf{X}$, it corresponds to a mention of type k that spans all words from position i to j inclusive. The resulting combination of mentions is unique. □

1.3 Inference

Our inference algorithm aims at solving two problems: computing the partition function and searching for the most probable hyperpath. It can be viewed as a generalized inside-outside style message-passing algorithm.

Specifically, to compute the following partition function, we need to sum over all the possible hyperpath.

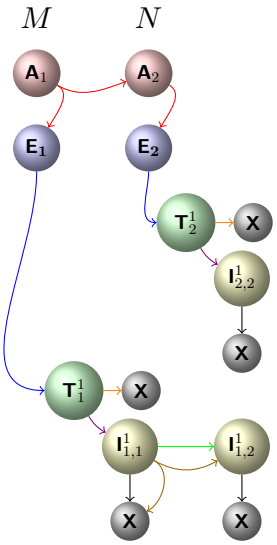

Figure 4: A complete segmental graph for a two-word sentence with one possible mention type.

$$\mathcal{Z} = \sum_{\boldsymbol{y}'} \exp f(\boldsymbol{x}, \boldsymbol{y}') = \sum_{\boldsymbol{y}'} \exp[\sum_{e \in \mathcal{G}_{\boldsymbol{y}'}} \psi(e, \boldsymbol{x})] \tag{1}$$

The summation can be decomposed into propagation from leaf nodes $\mathbf{X}$ to the root node $\mathbf{A}_1$ recursively. Let us show an example of a simple segmental hypergraph in Figure 4 to see how the algorithm works.

In this complete segmental hypergraph, there exist 8 kinds of hyperpath, each corresponds to a mention combination. We list the scores for each combination as follows.

- $\mathcal{G}_1$: mention M, N, MN

$$\begin{aligned} & f(\boldsymbol{x}, \mathcal{G}_1) = \\ & \psi(\{\mathbf{A}_1 \rightarrow (\mathbf{A}_2, \mathbf{E}_1)\}, \boldsymbol{x}) + \psi(\{\mathbf{A}_2 \rightarrow \mathbf{E}_2\}, \boldsymbol{x}) \\ & + \psi(\{\mathbf{E}_1 \rightarrow \mathbf{T}^1_1\}, \boldsymbol{x}) + \psi(\{\mathbf{E}_2 \rightarrow \mathbf{T}^1_2\}, \boldsymbol{x}) \\ & + \psi(\{\mathbf{T}^1_1 \rightarrow \mathbf{I}^1_{1,1}\}, \boldsymbol{x}) + \psi(\{\mathbf{I}^1_{1,1} \rightarrow (\mathbf{I}^1_{1,2}, \mathbf{X})\}, \boldsymbol{x}) \\ & + \psi(\{\mathbf{I}^1_{1,2} \rightarrow \mathbf{X}\}, \boldsymbol{x}) + \psi(\{\mathbf{T}^1_2 \rightarrow \mathbf{I}^1_{2,2}\}, \boldsymbol{x}) \\ & + \psi(\{\mathbf{I}^1_{2,2} \rightarrow \mathbf{X}\}, \boldsymbol{x}) \end{aligned} \tag{2}$$

- $\mathcal{G}_2$: mention MN, N

$$
\begin{aligned}
&f(\boldsymbol{x}, \mathcal{G}_2) = \\
&\psi(\{\mathbf{A}_1 \to (\mathbf{A}_2, \mathbf{E}_1)\}, \boldsymbol{x}) + \psi(\{\mathbf{A}_2 \to \mathbf{E}_2\}, \boldsymbol{x}) \\
&+ \psi(\{\mathbf{E}_1 \to \mathbf{T}_1^1\}, \boldsymbol{x}) + \psi(\{\mathbf{E}_2 \to \mathbf{T}_2^1\}, \boldsymbol{x}) \\
&+ \psi(\{\mathbf{T}_1^1 \to \mathbf{I}_{1,1}^1\}, \boldsymbol{x}) + \psi(\{\mathbf{I}_{1,1}^1 \to \mathbf{I}_{1,2}^1\}, \boldsymbol{x}) \\
&+ \psi(\{\mathbf{I}_{1,2}^1 \to \mathbf{X}\}, \boldsymbol{x}) + \psi(\{\mathbf{T}_2^1 \to \mathbf{I}_{2,2}^1\}, \boldsymbol{x}) \\
&+ \psi(\{\mathbf{I}_{2,2}^1 \to \mathbf{X}\}, \boldsymbol{x})
\end{aligned} \tag{3}
$$

- $\mathcal{G}_3$: mention M, N

$$
\begin{aligned}
&f(\boldsymbol{x}, \mathcal{G}_3) = \\
&\psi(\{\mathbf{A}_1 \to (\mathbf{A}_2, \mathbf{E}_1)\}, \boldsymbol{x}) + \psi(\{\mathbf{A}_2 \to \mathbf{E}_2\}, \boldsymbol{x}) \\
&+ \psi(\{\mathbf{E}_1 \to \mathbf{T}_1^1\}, \boldsymbol{x}) + \psi(\{\mathbf{E}_2 \to \mathbf{T}_2^1\}, \boldsymbol{x}) \\
&+ \psi(\{\mathbf{T}_1^1 \to \mathbf{I}_{1,1}^1\}, \boldsymbol{x}) + \psi(\{\mathbf{I}_{1,1}^1 \to \mathbf{X}\}, \boldsymbol{x}) \\
&+ \psi(\{\mathbf{T}_2^1 \to \mathbf{I}_{2,2}^1\}, \boldsymbol{x}) + \psi(\{\mathbf{I}_{2,2}^1 \to \mathbf{X}\}, \boldsymbol{x})
\end{aligned} \tag{4}
$$

- $\mathcal{G}_4$: mention N

$$
\begin{aligned}
&f(\boldsymbol{x}, \mathcal{G}_4) = \\
&\psi(\{\mathbf{A}_1 \to (\mathbf{A}_2, \mathbf{E}_1)\}, \boldsymbol{x}) + \psi(\{\mathbf{A}_2 \to \mathbf{E}_2\}, \boldsymbol{x}) \\
&+ \psi(\{\mathbf{E}_1 \to \mathbf{T}_1^1\}, \boldsymbol{x}) + \psi(\{\mathbf{E}_2 \to \mathbf{T}_2^1\}, \boldsymbol{x}) \\
&+ \psi(\{\mathbf{T}_1^1 \to \mathbf{X}\}, \boldsymbol{x}) + \psi(\{\mathbf{T}_2^1 \to \mathbf{I}_{2,2}^1\}, \boldsymbol{x}) \\
&+ \psi(\{\mathbf{I}_{2,2}^1 \to \mathbf{X}\}, \boldsymbol{x})
\end{aligned} \tag{5}
$$

- $\mathcal{G}_5$: mention M, MN

$$
\begin{aligned}
&f(\boldsymbol{x}, \mathcal{G}_5) = \\
&\psi(\{\mathbf{A}_1 \to (\mathbf{A}_2, \mathbf{E}_1)\}, \boldsymbol{x}) + \psi(\{\mathbf{A}_2 \to \mathbf{E}_2\}, \boldsymbol{x}) \\
&+ \psi(\{\mathbf{E}_1 \to \mathbf{T}_1^1\}, \boldsymbol{x}) + \psi(\{\mathbf{E}_2 \to \mathbf{T}_2^1\}, \boldsymbol{x}) \\
&+ \psi(\{\mathbf{T}_1^1 \to \mathbf{I}_{1,1}^1\}, \boldsymbol{x}) \\
&+ \psi(\{\mathbf{I}_{1,1}^1 \to (\mathbf{I}_{1,2}^1, \mathbf{X})\}, \boldsymbol{x}) \\
&+ \psi(\{\mathbf{I}_{1,2}^1 \to \mathbf{X}\}, \boldsymbol{x}) + \psi(\{\mathbf{T}_2^1 \to \mathbf{X}\}, \boldsymbol{x})
\end{aligned} \tag{6}
$$

- $\mathcal{G}_6$: mention MN

$$
\begin{aligned}
&f(\boldsymbol{x}, \mathcal{G}_6) = \\
&\psi(\{\mathbf{A}_1 \to (\mathbf{A}_2, \mathbf{E}_1)\}, \boldsymbol{x}) + \psi(\{\mathbf{A}_2 \to \mathbf{E}_2\}, \boldsymbol{x}) \\
&+ \psi(\{\mathbf{E}_1 \to \mathbf{T}_1^1\}, \boldsymbol{x}) + \psi(\{\mathbf{E}_2 \to \mathbf{T}_2^1\}, \boldsymbol{x}) \\
&+ \psi(\{\mathbf{T}_1^1 \to \mathbf{I}_{1,1}^1\}, \boldsymbol{x}) + \psi(\{\mathbf{I}_{1,1}^1 \to \mathbf{I}_{1,2}^1\}, \boldsymbol{x}) \\
&+ \psi(\{\mathbf{I}_{1,2}^1 \to \mathbf{X}\}, \boldsymbol{x}) + \psi(\{\mathbf{T}_2^1 \to \mathbf{X}\}, \boldsymbol{x})
\end{aligned} \tag{7}
$$

- $\mathcal{G}_7$: mention M

$$
\begin{aligned}
&f(\boldsymbol{x}, \mathcal{G}_7) = \\
&\psi(\{\mathbf{A}_1 \to (\mathbf{A}_2, \mathbf{E}_1)\}, \boldsymbol{x}) + \psi(\{\mathbf{A}_2 \to \mathbf{E}_2\}, \boldsymbol{x}) \\
&+ \psi(\{\mathbf{E}_1 \to \mathbf{T}_1^1\}, \boldsymbol{x}) + \psi(\{\mathbf{E}_2 \to \mathbf{T}_2^1\}, \boldsymbol{x}) \\
&+ \psi(\{\mathbf{T}_1^1 \to \mathbf{I}_{1,1}^1\}, \boldsymbol{x}) + \psi(\{\mathbf{I}_{1,1}^1 \to \mathbf{X}\}, \boldsymbol{x}) \\
&+ \psi(\{\mathbf{T}_2^1 \to \mathbf{X}\}, \boldsymbol{x})
\end{aligned} \tag{8}
$$

- $\mathcal{G}_8$: no mention

$$
\begin{aligned}
&f(\boldsymbol{x}, \mathcal{G}_8) = \\
&\psi(\{\mathbf{A}_1 \to (\mathbf{A}_2, \mathbf{E}_1)\}, \boldsymbol{x}) + \psi(\{\mathbf{A}_2 \to \mathbf{E}_2\}, \boldsymbol{x}) \\
&+ \psi(\{\mathbf{E}_1 \to \mathbf{T}_1^1\}, \boldsymbol{x}) + \psi(\{\mathbf{E}_2 \to \mathbf{T}_2^1\}, \boldsymbol{x}) \\
&+ \psi(\{\mathbf{T}_1^1 \to \mathbf{X}\}, \boldsymbol{x}) + \psi(\{\mathbf{T}_2^1 \to \mathbf{X}\}, \boldsymbol{x})
\end{aligned} \tag{9}
$$

The partition function of this example can be written as follows:

$$
\mathcal{Z} = \sum_{i=1}^{8} \exp f(\boldsymbol{x}, \mathcal{G}_i) = \sum_{i=1}^{8} \exp[\sum_{e \in \mathcal{G}_i} \psi(e, \boldsymbol{x})] \tag{10}
$$

Though the number of hyperpath is exponential in the number of words, the summation over them can be optimized through dynamic programming based on the observation that hyperpaths share common sub-structures. For example, $\mathcal{G}_1$, $\mathcal{G}_2$ and $\mathcal{G}_3$ only differs with the sub-structure from $\mathbf{I}_{1,1}^1$ to leaf nodes. So the sum of them can be written as:

$$
\begin{aligned}
&\exp\big(f(\boldsymbol{x}, \mathcal{G}_1) + f(\boldsymbol{x}, \mathcal{G}_2) + f(\boldsymbol{x}, \mathcal{G}_3)\big) = \\
&\quad \exp\big(\psi(\{\mathbf{A}_1 \to (\mathbf{A}_2, \mathbf{E}_1)\}, \boldsymbol{x}) + \psi(\{\mathbf{A}_2 \to \mathbf{E}_2\}, \boldsymbol{x}) \\
&\quad + \psi(\{\mathbf{E}_1 \to \mathbf{T}_1^1\}, \boldsymbol{x}) + \psi(\{\mathbf{E}_2 \to \mathbf{T}_2^1\}, \boldsymbol{x}) \\
&\quad + \psi(\{\mathbf{T}_1^1 \to \mathbf{I}_{1,1}^1\}, \boldsymbol{x})\big) \cdot \\
&\quad \Big(\exp\big(\psi(\{\mathbf{I}_{1,1}^1 \to (\mathbf{I}_{1,2}^1, \mathbf{X})\}, \boldsymbol{x}) + \psi(\{\mathbf{I}_{1,2}^1 \to \mathbf{X}\}, \boldsymbol{x}) \\
&\quad + \psi(\{\mathbf{T}_2^1 \to \mathbf{I}_{2,2}^1\}, \boldsymbol{x}) + \psi(\{\mathbf{I}_{2,2}^1 \to \mathbf{X}\}, \boldsymbol{x})\big) + \\
&\quad \exp\big(\psi(\{\mathbf{I}_{1,1}^1 \to \mathbf{I}_{1,2}^1\}, \boldsymbol{x}) + \psi(\{\mathbf{I}_{1,2}^1 \to \mathbf{X}\}, \boldsymbol{x}) \\
&\quad + \psi(\{\mathbf{T}_2^1 \to \mathbf{I}_{2,2}^1\}, \boldsymbol{x}) + \psi(\{\mathbf{I}_{2,2}^1 \to \mathbf{X}\}, \boldsymbol{x})\big) + \\
&\quad \exp\big(\psi(\{\mathbf{I}_{1,1}^1 \to \mathbf{X}\}, \boldsymbol{x}) + \psi(\{\mathbf{T}_2^1 \to \mathbf{I}_{2,2}^1\}, \boldsymbol{x}) \\
&\quad + \psi(\{\mathbf{I}_{2,2}^1 \to \mathbf{X}\}, \boldsymbol{x})\big)\Big)
\end{aligned} \tag{11}
$$

The second term can be reused when we need to compute $\exp\big(f(\boldsymbol{x}, \mathcal{G}_5) + f(\boldsymbol{x}, \mathcal{G}_6) + f(\boldsymbol{x}, \mathcal{G}_7)\big)$. In

	ACE-2004			ACE-2005			GENIA		
	Train (%)	Dev (%)	Test (%)	Train (%)	Dev (%)	Test (%)	Train (%)	Dev (%)	Test (%)
# sentences	6,799	829	879	7,336	958	1,047	14,836	1,855	1,855
with o.l.	2,683 (39)	293 (35)	272 (42)	2,683 (37)	340 (35)	330 (32)	3,199 (22)	366 (20)	448 (24)
# mentions	22,207	2,511	3,031	24,687	3,217	3,027	46,473	5,014	5,600
o.l.	10,170 (46)	1,091 (43)	1,418 (47)	9,937 (40)	1,192 (37)	1,184 (39)	8,337 (18)	915 (18)	1,217 (22)
o.l. (same type)	5,431 (24)	624 (25)	780 (26)	5,044 (20)	600 (19)	638 (21)	4,613 (10)	479 (10)	634 (11)
o.l. (same type & lb)	2,188 (10)	204 (8)	307 (10)	1,973 (8)	243 (8)	253 (8)	2,133 (5)	202 (4)	287 (5)
length > 6	1,439 (6)	179 (7)	199 (7)	1,343 (5)	148 (5)	160 (6)	2,449 (5)	302 (6)	301 (5)
max length	57	35	43	49	30	27	28	28	19

Table 1: Statistics for ACE-2004, ACE-2005 and GENIA. *o.l.*: overlapping mentions, *lb*: left boundary.

the case that the sentence length is greater than 2, the second term can also be reused for computation that involves previous $\mathbf{I}$ nodes.

Based on this intuition, we define a message function $\mu[\mathbf{p}]$ for each node $\mathbf{p}$, which can be viewed as the summation of sub-structures from $\mathbf{p}$ to leaf nodes. Let's set the message value for leaf nodes to be 0. Then our algorithm passes messages from leaf node to the root node $\mathbf{A}_1$ based on the following recursive computation.

$$\mu[\mathbf{p}] \leftarrow \log\Big(\sum_{e:h(e)\equiv\mathbf{p}} \exp\big(\psi(e,\boldsymbol{x}) + \sum_{\mathbf{c}\in\mathcal{T}(e)} \mu[\mathbf{c}]\big)\Big) \quad (12)$$

where $h(e)$ is the head of the hyperedge e, and $\mathcal{T}(e)$ is the collection of nodes that form the tail of e – they are the child nodes of $h(e)$ given e.

For example,

$$\mu[\mathbf{I}^1_{1,2}] \leftarrow \log\Big(\exp\big(\psi(\{\mathbf{I}^1_{1,2} \rightarrow \mathbf{X}\}, \boldsymbol{x}) + \mu[\mathbf{X}]\big)\Big) \quad (13)$$

$$\begin{aligned}\mu[\mathbf{I}^1_{1,1}] \leftarrow \log\Big(&\exp\big(\psi(\{\mathbf{I}^1_{1,1} \rightarrow \mathbf{I}^1_{1,2}\}, \boldsymbol{x}) + \mu[\mathbf{I}^1_{1,2}]\big) \\ &+ \exp\big(\psi(\{\mathbf{I}^1_{1,1} \rightarrow \mathbf{X}\}, \boldsymbol{x}) + \mu[\mathbf{X}]\big) \\ &+ \exp\big(\psi(\{\mathbf{I}^1_{1,1} \rightarrow \{\mathbf{I}^1_{1,2}\}, \mathbf{X}\}\}, \boldsymbol{x}) + \mu[\mathbf{X}] + \mu[\mathbf{I}^1_{1,2}]\big)\Big)\end{aligned} \quad (14)$$

Ultimately, the value of partition function $\mathcal{Z}$ is equal to the message value of the root node which is $\mu[\mathbf{A}_1]$.

For MAP inference, we need to search for the hyperpath with the most scores. We only need to replace the sum operation in (12) to max operation.

$$\mu[\mathbf{p}] \leftarrow \max_{e:h(e)\equiv\mathbf{p}} \Big(\phi(e,\boldsymbol{x}) + \sum_{\mathbf{c}\in\mathcal{T}(e)} \mu[\mathbf{c}]\Big) \quad (15)$$

The $\mu[\mathbf{p}]$ can be understood as the maximal scores for all sub-structures from $\mathbf{p}$ to leaf nodes. Analogously, $\mu[\mathbf{A}_1]$ stores the score for the best hyperpath, which can be retrieved using backtracking.

2 Experiments

This section provides the complete statistics for ACE and GENIA datasets, experiment on length restriction, the handcrafted features used for baseline models, a list of hyperparameters used for training, some complete results of experiments in the paper.

2.1 Statistics

Table 1 shows the complete statistics for ACE datasets and GENIA.

2.2 Handcrafted Features

The non-neural baselines methods all use the same set of features excluding the results reported from (Finkel and Manning, 2009) who design their own handcrafted features. These features include word-level features such as surrounding words (and POS tags), word n-grams, bag-of-words and word patterns [1]. For span-level features, we use capitalization patterns inside a span and indicators for bigrams and trigrams in a span, as well as capitalization patterns in 3-word windows before and after the span, inspired by (Sarawagi and Cohen, 2005). In GENIA dataset, we also include Brown cluster features learned on PubMed abstracts, following (Finkel and Manning, 2009).

For the neural model of FOFE, it was trained using the same embeddings as our neural segmental hypergraph.

2.3 Hyperparameters

Table 3 lists the hyperparameters that were used for training our neural segmental hypergraph in ACE and GENIA datasets.

2.4 Effect of Length Restriction

We enumerate the possible lengths between [4, 20] to see how it affect the performance. Experiments are conducted in ACE04 and GENIA. As shown in

[1] Full descriptions of features can be found in their original paper (Lu and Roth, 2015)

	ACE-2004 (TEST)							ACE-2005 (TEST)						
	Overlapping			Non-Overlapping			*w/s*	Overlapping			Non-Overlapping			*w/s*
	P	R	F_1	P	R	F_1		P	R	F_1	P	R	F_1	
Lu and Roth (2015)	72.5	52.4	60.8	72.5	65.0	68.6	460	68.1	52.6	59.4	64.1	65.1	64.6	503
Muis and Lu (2017)	72.1	55.3	62.6	74.1	65.5	69.5	251	70.4	55.0	61.8	67.2	63.4	65.2	253
SH (c=6)	78.2	65.6	71.3	80.0	68.0	73.5	263	80.2	68.3	73.8	74.8	70.0	72.3	248
SH (c=n)	77.3	72.2	74.7	77.1	70.9	73.8	171	80.6	73.6	76.9	75.5	71.5	73.4	157

Table 2: Detailed results on ACE04 and ACE05, *w/s*: number of words decoded per second.

Hyperparameter	ACE2004	ACE2005	GENIA
word embedding dim	100	100	100
pos embeddings dim	32	32	32
LSTM(word) hidden size	128	128	256
LSTM(span) hidden size	128	128	256
dropout	0.5	0.5	0.6
softmax margin β	2.5	2.5	3
l_2	0.0001	0.0001	0.0001

Table 3: Hyperparameters of training neural segmental hypergraph. dim: dimension

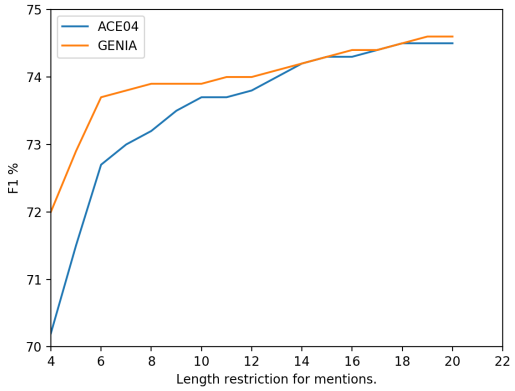

Figure 5: Effect of length restrictions

Figure 5, the performance improves consistently as we increase the maximal length allowed for a mention. Moreover, the effect becomes less significant as the length grows larger since the number of long mentions that are additionally recalled also decreases. Note that larger length could also introduce more false positives, but our system can eliminate such negative effect, revealing the robustness of our model.

2.5 Handling Overlapping Mentions

Table 2 shows the complete results of handling overlapping mentions in both ACE04 and ACE05 datasets. Our neural segmental hypergraph yields similar performance in these two datasets.

2.6 Detailed Results on CoNLL 2003

We made two sets of comparisons. The results can be found in Table 4. First, we compare our models

Model	**F_1**
CRF (LINEAR)	83.8
CRF (CASCADED)	84.3
Semi-CRF (c=6)	**85.3**
Semi-CRF (c=n)	84.9
Lu and Roth (2015)	83.5
Muis and Lu (2017)	84.3
SH (HAND-CRAFTED, c=6)	85.2
SH (HAND-CRAFTED, c=n)	84.6
SH (HAND-CRAFTED, SM, c=6)	**85.3**
SH (HAND-CRAFTED, SM, c=n)	84.8
SH (c=6)	89.6
SH (c=n)	89.2
SH (c=6) + *char*	90.5
SH (c=n) + *char*	90.2
Collobert et al. (2011)	88.7
Collobert et al. (2011) *	89.6
Huang et al. (2015) *	90.1
Chiu and Nichols (2016)	90.9
Chiu and Nichols (2016) *	**91.6**
Lample et al. (2016)	90.9
Ma and Hovy (2016)	91.2
Xu et al. (2017)	90.7
Strubell et al. (2017)	90.5

Table 4: CoNLL-2003 NER results on English portion. Models with * indicates that they're learned with external features excluding pre-trained embeddings.

with previous baselines with the same handcrafted features. We find that our segmental hypergraph achieves similar performance compared to the best model semi-CRF. The length restriction is beneficial in this dataset (unlike ACE and GENIA) since most mentions are very short.

In the second set, we compared our model with recent neural network based models. We also include some neural models that take advantage of external features in the list. Our neural segmental hypergraph achieves competitive results compared with other approaches and it could potentially be improved if external features are considered.

References

Jason PC Chiu and Eric Nichols. 2016. Named entity recognition with bidirectional lstm-cnns. *Transactions of the Association for Computational Linguistics* 4:357–370.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research* 12(Aug):2493–2537.

Jenny Rose Finkel and Christopher D Manning. 2009. Nested named entity recognition. In *Proc. of EMNLP*.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991* .

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proc. of NAACL-HLT*.

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*.

Aldrian Obaja Muis and Wei Lu. 2017. Labeling gaps between words: Recognizing overlapping mentions with mention separators. In *Proc. of EMNLP*.

Sunita Sarawagi and William W Cohen. 2005. Semimarkov conditional random fields for information extraction. In *Proc. of NIPS*.

Emma Strubell, Patrick Verga, David Belanger, and Andrew McCallum. 2017. Fast and accurate entity recognition with iterated dilated convolutions. In *Proc. of EMNLP*.

Bailin Wang and Wei Lu. 2018. Neural segmental hypergraphs for overlapping mention recognition. In *EMNLP*.

Mingbin Xu, Hui Jiang, and Sedtawut Watcharawittayakul. 2017. A local detection approach for named entity recognition and mention detection. In *Proc. of ACL*.