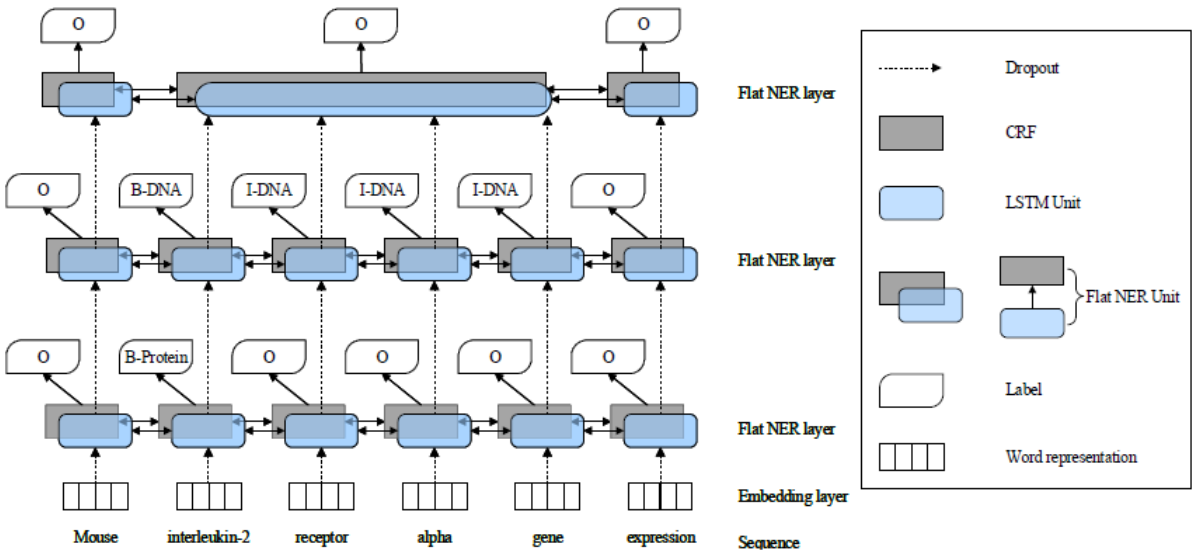
A Boundary-aware Neural Model for Nested Named Entity Recognition

Changmeng Zheng[1], **Yi Cai**[1]*, **Jingyun Xu**[1], **Ho-fung Leung**[2] and **Guandong Xu**[3]
[1]School of Software Engineering, South China University of Technology, Guangzhou, China
[2]The Chinese University of Hong Kong, Hong Kong SAR, China
[3]Advanced Analytics Institute, University of Technology Sydney, Australia
sethecharm@mail.scut.edu.cn, ycai@scut.edu.cn

Recent

- Layered sequence labeling model : first extract the inner entities (contained by other entities) and feed them into the next layer to extract outer entities → error propagation
- Exhaustive region classification model : enumerates all possible regions or spans in sentences to predict entities in a single layer → extraction of some non-entities



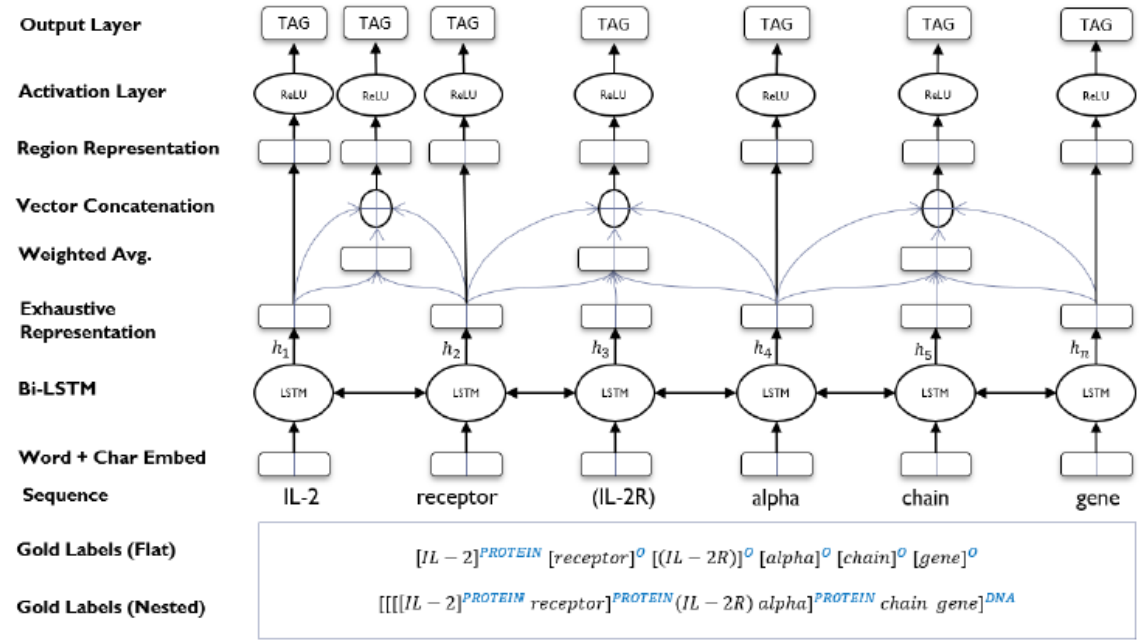
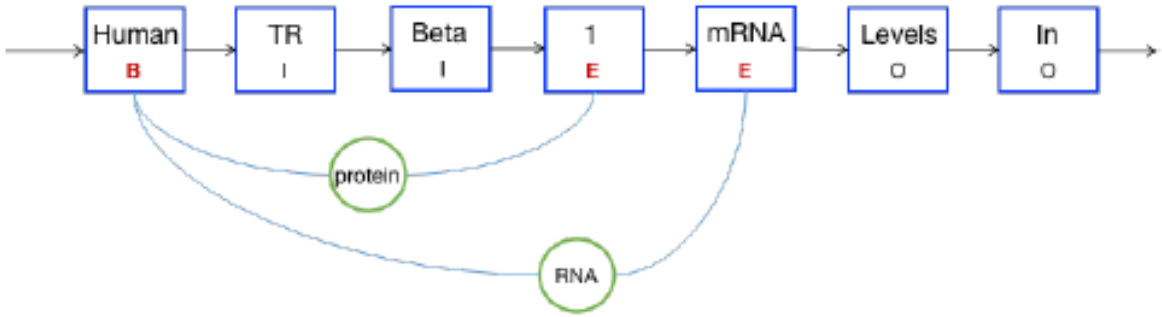


Figure 1: Architecture of the proposed neural exhaustive model. The model considers all possible regions up to a maximum size, but we depict here only a small subset for brevity. "*IL-2*", "*IL-2 receptor*", "*IL-2 receptor (IL-2R) alpha*", and "*IL-2 receptor (IL-2R) alpha chain gene*" are nested entities.

Summary

- propose **a boundary-aware neural model** which leverages entity boundaries to predict categorical labels. Firstly it can locate entities precisely, and utilizes boundary-relevant regions to predict entity categorical labels.
- multitask learning to capture the dependencies of entity boundaries and their categorical labels



recently	O	O	O	O
circulating	O	B-cell_type	O	O
human	B-cell_type	I-cell_type	O	O
lymphocytes	I-cell_type	I-cell_type	O	O
have	O	O	O	O
provided	O	O	O	O
increasing	O	O	O	O
evidence	O	O	O	O

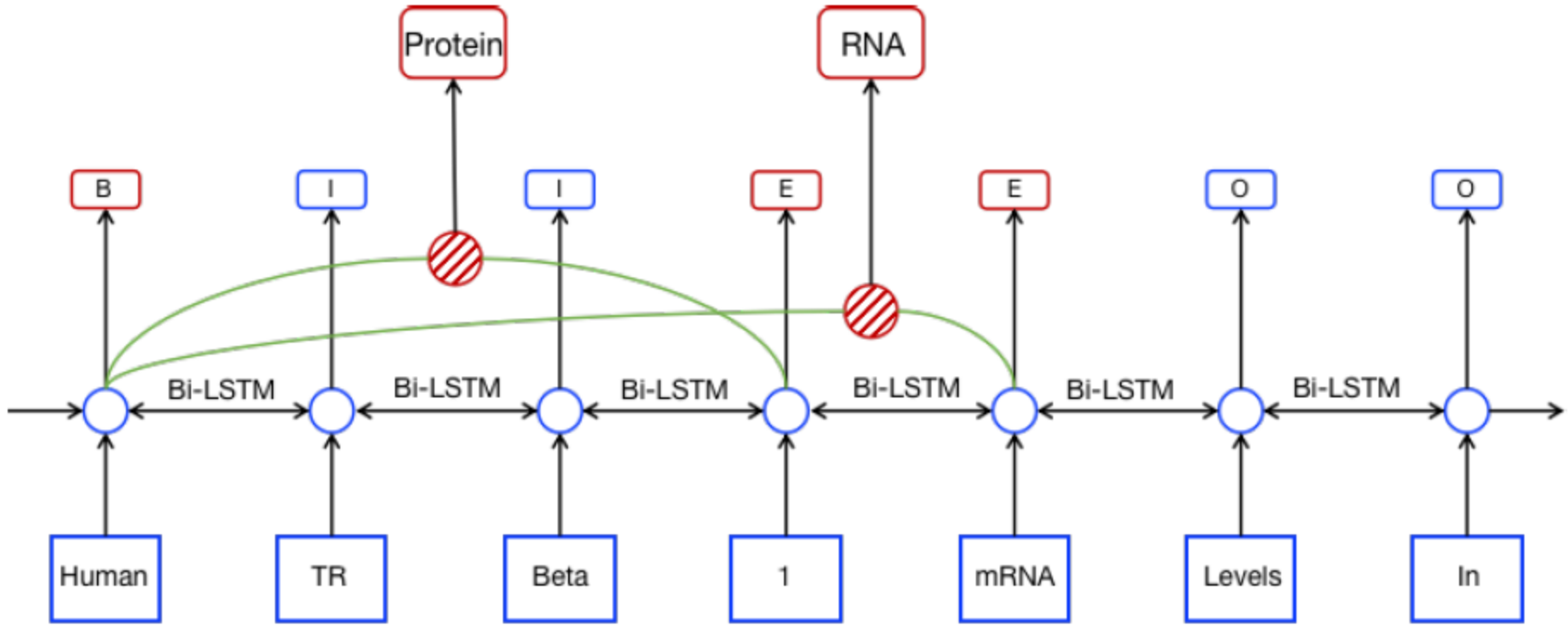


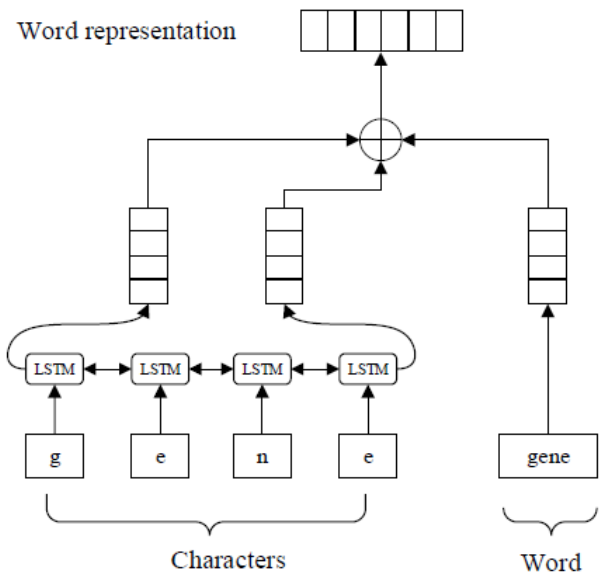
Figure 2: The Architecture of our boundary-aware model. The representation of each token in sentence "Human TR Beta 1 mRNA Levels in.." is feed into a shared bidirectional LSTM layer. We leverage the outputs of Bi-LSTM to detect entity boundaries and their categorical labels. The red circle indicates entity region representations between entity boundaries.

3.1 Token Representation

For a given sentence consisting of n tokens (t_1,t_2,....t_n), we represent the word embedding of i-th token t_i as equation(1):

$$\mathbf{x}_i^w = \mathbf{e}^w(t_i) \tag{1}$$

where $\mathbf{e}^w$ denotes a word embedding lookup table.



sentations. Denoting the representation of characters within t_i as $\mathbf{x}_i^c$, The embedding of each character within token t_i is denoted as $\mathbf{e}^c(c_j)$. $\mathbf{e}^c$ is the character embedding lookup which is initialized randomly. Then we feed them into a bidirectional LSTM layer to learn hidden states. The forward and backward outputs are concatenated to construct character representations:

$$\mathbf{x}_i^c = [\overrightarrow{\mathbf{h}_i^c}; \overleftarrow{\mathbf{h}_i^c}] \tag{2}$$

$$\mathbf{x}_i^t = [\mathbf{x}_i^w; \mathbf{x}_i^c]$$

3.2 Shared Feature Extractor

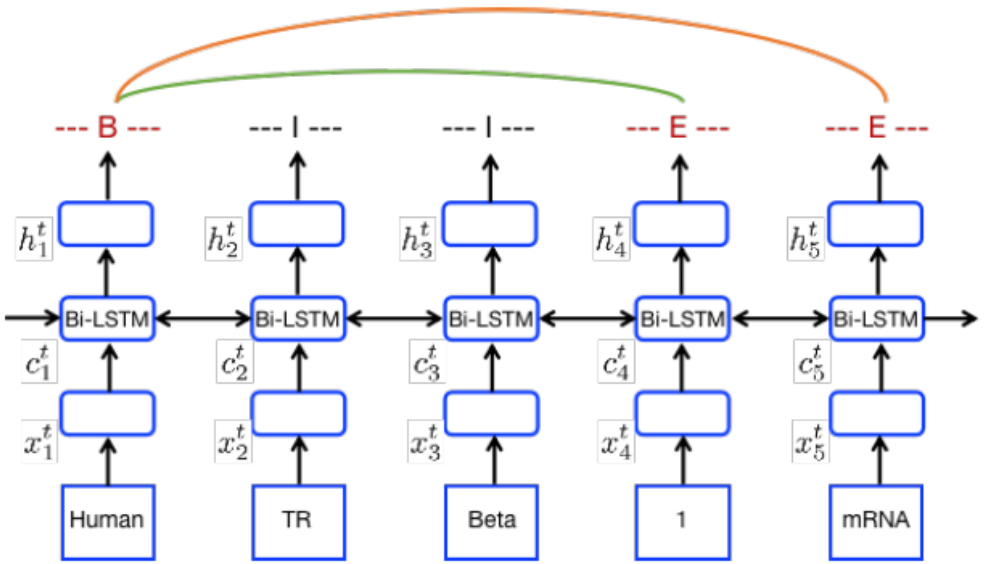
$$\overrightarrow{\mathbf{h}}_i^t = \overrightarrow{\mathrm{LSTM}}(\mathbf{x}_i^t, \overrightarrow{\mathbf{h}}_{i-1}^t) \quad (4)$$

$$\overleftarrow{\mathbf{h}}_i^t = \overleftarrow{\mathrm{LSTM}}(\mathbf{x}_i^t, \overleftarrow{\mathbf{h}}_{i-1}^t) \quad (5)$$

$$\mathbf{h}_i^t = [\overrightarrow{\mathbf{h}}_i^t; \overleftarrow{\mathbf{h}}_i^t] \quad (6)$$

3.3 Entity Boundary Detection

boundary labels first. Formally, given a sentence (t_1,t_2,...t_n), and one entity in the sentence. we represent the entity as $R(i, j)$, which denotes the entity is composed by a continuous token sequence (t_i,t_{i+1},...t_j). Specially, we tag the boundary token t_i as "B" and t_j as "E". The tokens inside entities are assigned with label "I" and non-entity tokens are assigned with "O" labels.



$$\mathbf{o}_i^t = \mathbf{U}\mathbf{h}_i^t + \mathbf{b}$$

$$\mathbf{d}_i^t = \mathrm{softmax}(\mathbf{o}_i^t)$$

$$L_{bcls} = -\sum(\hat{\mathbf{d}}_i^t)\log(\mathbf{d}_i^t)$$

3.4 Entity Categorical Label Prediction

$$\mathrm{R}_{i,j} = \left[\frac{1}{j-i+1} \sum_{k=i}^{j} \mathbf{h}_k^t \right]$$

$$\mathbf{d}_{i,j}^e = \mathrm{softmax}(\mathbf{U}_{i,j}^e \mathrm{R}_{i,j} + \mathbf{b}_{i,j}^e)$$

$$L_{ecls} = -\sum (\hat{\mathbf{d}}_{i,j}^e) \log(\mathbf{d}_{i,j}^e)$$

3.5 Multitask Training

In our model, it is inconvenient and inefficient for the reason that we predict entity categorical labels after all boundary-relative regions have been detected. Considering our boundary detection module and entity categorical label prediction module share the same entity boundaries, we apply a multitask loss for training the two tasks simultaneously.

During training phase, we feed the ground-truth boundary labels into entity categorical label prediction module so that the classifier will be trained without affection from error boundary detection. As for testing phase, the outputs of boundary detection will be collected. The detected boundaries will indicate which entity regions should be considered into predicting categorical labels. The multitask loss function is defined as follows:

$$L_{multi} = \alpha \sum L_{bcls} + (1-\alpha) \sum L_{ecls} \quad (13)$$

Experiment

- GENIA (Kim et al.,2003)
- JNLPBA (Kim et al., 2004)
- GermEval 2014 (Benikova et al., 2014)

Item	Train	Dev	Test	Overall	Nested
Document	1599	189	212	2000	-
Sentences	15023	1669	1854	18546	-
Percentage	81%	9%	10%	100%	-
DNA	7650	1026	1257	9933	1744
RNA	692	132	109	933	407
Protein	28728	2303	3066	34097	1902
Cell Line	3027	325	438	3790	347
Cell Type	5832	551	604	6987	389
Overall	45929	4337	5474	55740	4789

Model	P(%)	R(%)	F(%)
Finkel and Manning (2009)[3]	75.4	65.9	70.3
Lu and Roth (2015)[3]	72.5	65.2	68.7
Muis and Lu (2017)[3]	75.4	66.8	70.8
Sohrab and Miwa (2018)	73.3	68.3	70.7
Ju et al. (2018)	**76.1**	66.8	71.1
Our model(softmax)	75.9	**73.6**	**74.7**
Our model(CRF)	74.6	73.2	73.9

Table 2: Performance on GENIA test set. Our models with softmax and CRF outperform other state-of-the-art methods.

Model	P(%)	R(%)	F(%)
Sohrab and Miwa (2018)	**75.0**	60.8	67.2
Ju et al. (2018)	72.9	61.5	66.7
Our model	74.5	**69.1**	**71.7**

Table 3: Performance on GermEval 2014 test set. Our model outperforms two state-of-the-art methods in nested NER.

Category	P(%)	R(%)	F(%)	Ju. F(%)	Soh. F(%)
DNA	73.6	67.8	**70.6**	70.1	67.8
RNA	82.2	80.7	**81.5**	80.8	75.9
protein	76.7	76.0	**76.4**	72.7	72.9
cell line	77.8	65.8	**71.3**	66.9	63.6
cell type	73.9	71.2	**72.5**	71.3	69.8
overall	75.8	73.6	**74.7**	71.1	70.7

Table 4: Our results on five categories compared to Ju et al. (2018) and Sohrab and Miwa (2018) on GENIA test set.

Model	Boundary Detection		
	P(%)	R(%)	F(%)
Sohrab and Miwa (2018)	76.6	69.2	72.7
Ju et al. (2018)	79.9	67.08	73.4
Our model(softmax)	79.7	**76.9**	**78.3**

Table 5: Performance of Boundary Detection on GENIA test set.

Boundary Label	P(%)	R(%)	F(%)
O (non-entity)	99.3	99.0	99.2
B (beginning)	84.4	84.3	84.3
E (end)	86.0	87.2	86.6
I (inner-entity)	82.8	88.6	85.6

Table 6: Performance of Boundary Label Prediction with softmax classifier on GENIA test set.

Sentence	Cloning of a transcriptionally active human TATA binding factor.
Gold Label	protein: {human TATA binding factor; transcriptionally active human TATA binding factor}
Exhaustive model	protein: {TATA binding factor; transcriptionally active human TATA binding factor}
Layered model	protein: {transcriptionally active human TATA binding factor}
Our model(pipeline)	protein: {human TATA binding factor;}
Our model(multitask)	protein: {human TATA binding factor; transcriptionally active human TATA binding factor}

Table 9: An example of predicted results in GENIA test dataset.