Nested Named Entity Recognition

2020/04/21

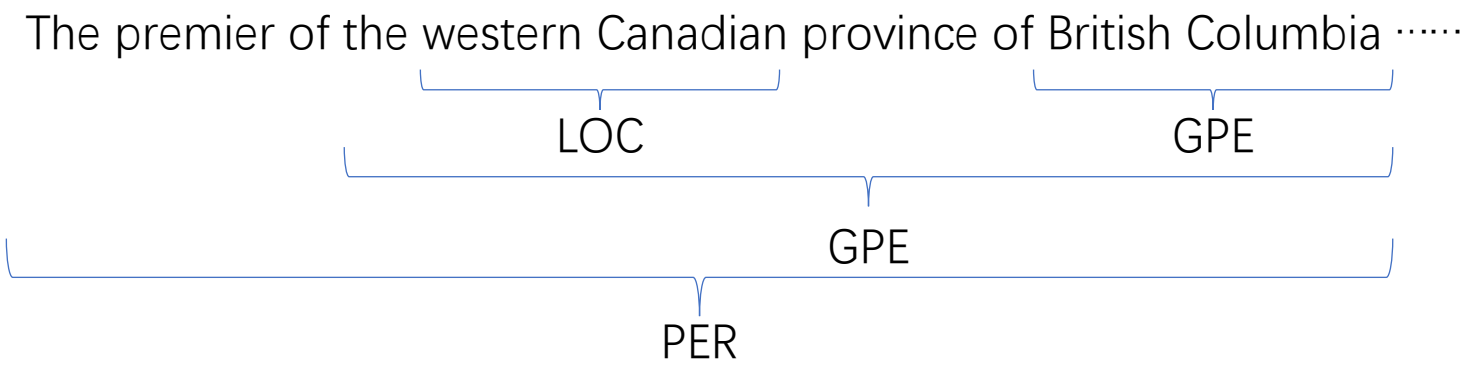
CQU 1701 Online Meeting

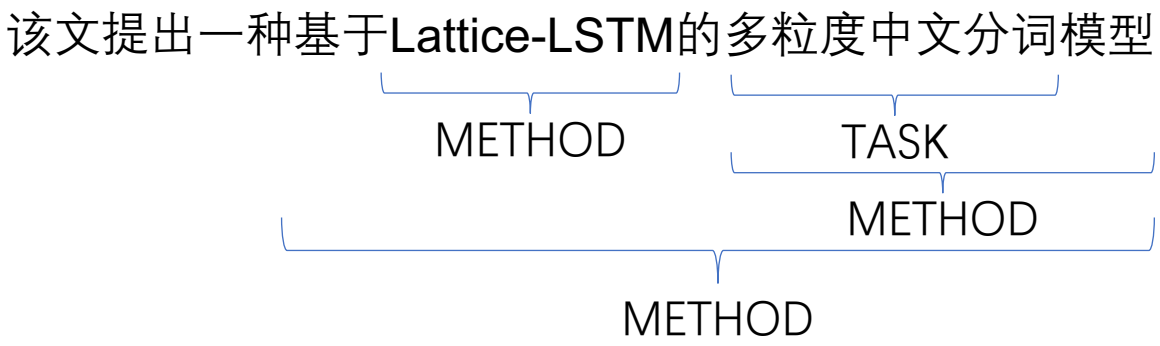
Contents

- Problem Definition
- Datasets
- Sequence Labeling Methods
- Sub-Sequence Classifier Methods
- Summary

Problem Definition

Embedded names which are included in other entities





Problem Definition

- Give a sequence $X = \{x_1, x_2, \dots\dots, x_n\}$
- Predict a sequence of label $Y = \{y_1, y_2, \dots\dots, y_n\}$
- y_n contains a list of labels not a single label:

$$y_n = \{y_n^1, y_n^2, \dots, y_n^m\}, \qquad m = nested\_layer$$

A multi-label classification problem

Nested Type

- A nested entity contains more than 1 entity:
  - the western Canadian province of British Columbia
- A nested entity extended from 1 flat entity + key words:
  - 多粒度中文分词模型

Datasets

- GENIA
- ACE2005
- NNE

	Text Type	Doc Number	Mentions	Entity Types	Nested Level
GENIA	Biomedical Text	2,000	92,681	36	4
ACE2005	News	464	30,966	7	6
NNE	News	2,312	279,795	114	6

Raw & Processed format

▷Mice◁T5 transgenic for the ▷▷human T cell leukemia virus◁T7 (▷HTLV-I◁T8) ▷Tax◁T9 gene◁T6 develop ▷fibroblastic tumors◁T10 that express ▷NF-kappa B-inducible early genes◁T11.

<sentence id="S2"><term id="T5" sem="Multicellular_organism">Mice</term> transgenic for the <term id="T6" sem="DNA_domain_or_region"><term id="T7" sem="Virus">human T cell leukemia virus</term> (<term id="T8" sem="Virus">HTLV-I</term>) <term id="T9" sem="Protein_molecule">Tax</term> gene</term> develop <term id="T10" sem="Tissue">fibroblastic tumors</term> that express <term id="T11" sem="DNA_family_or_group">NF-kappa B-inducible early genes</term>.</sentence>

```
These   O   O   O   O
data    O   O   O   O
indicate    O   O   O   O
that    O   O   O   O
IL  B-protein   O   O   O
-   I-protein   O   O   O
4   I-protein   O   O   O
suppresses  O   O   O   O
the O   O   O   O
induction   O   O   O   O
of  O   O   O   O
transcription   B-protein   O   O   O
factors I-protein   O   O   O
in  O   O   O   O
human   O   B-cell_type O   O
activated   O   I-cell_type O   O
monocytes   B-cell_type I-cell_type O   O
.   O   O   O   O
```

Sequence Labeling Methods

- Combined Label
- Neural Layered Model

BILOU Encoding(Combined Label)

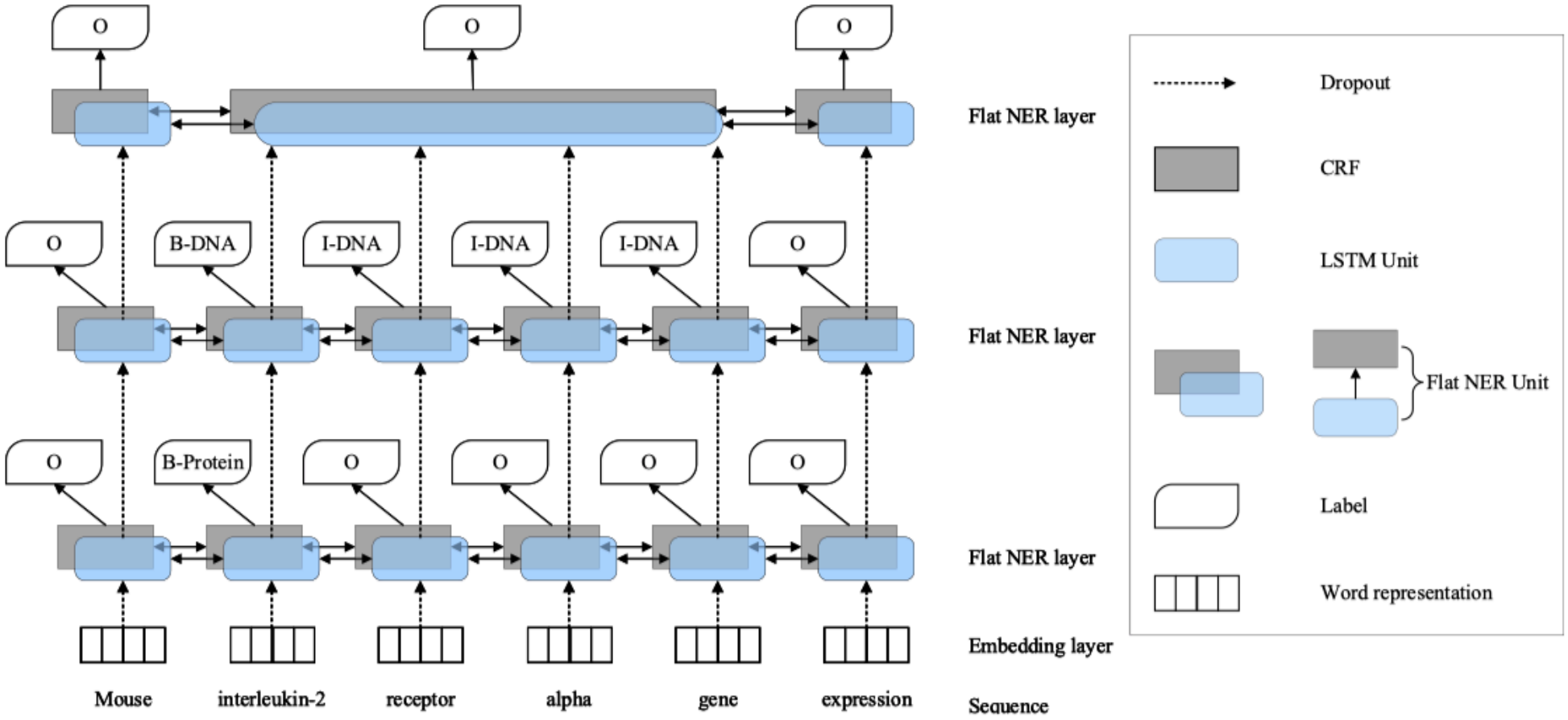
- Flat NER:
  - Begin, Inside, Outside
  - Type1, Type2, Type3

  } B-Type1, I-Type1,……
- Nested NER:

in	O
the	B-ORG
US	I-ORG|U-GPE
Federal	I-ORG
District	I-ORG|U-GPE
Court	I-ORG
of	I-ORG
New	I-ORG|B-GPE
Mexico	L-ORG|L-GPE
.	O

Neural Architectures for Nested NER through Linearization, ACL 2019

Neural Layered Model



A Neural Layered Model for Nested Named Entity Recognition, NAACL-HLT 2018

Problems

- Combined Label:
  - The number of Labels grows exponentially
  - Label distribution is too sparse
- Layered Model:
  - Error propagation
  - Can not train model parallelly

Sub-Sequence Classifier Methods

- Deep Exhaustive Model
- Boundary-aware Neural Model
- Connection-aware Model

Simple Idea

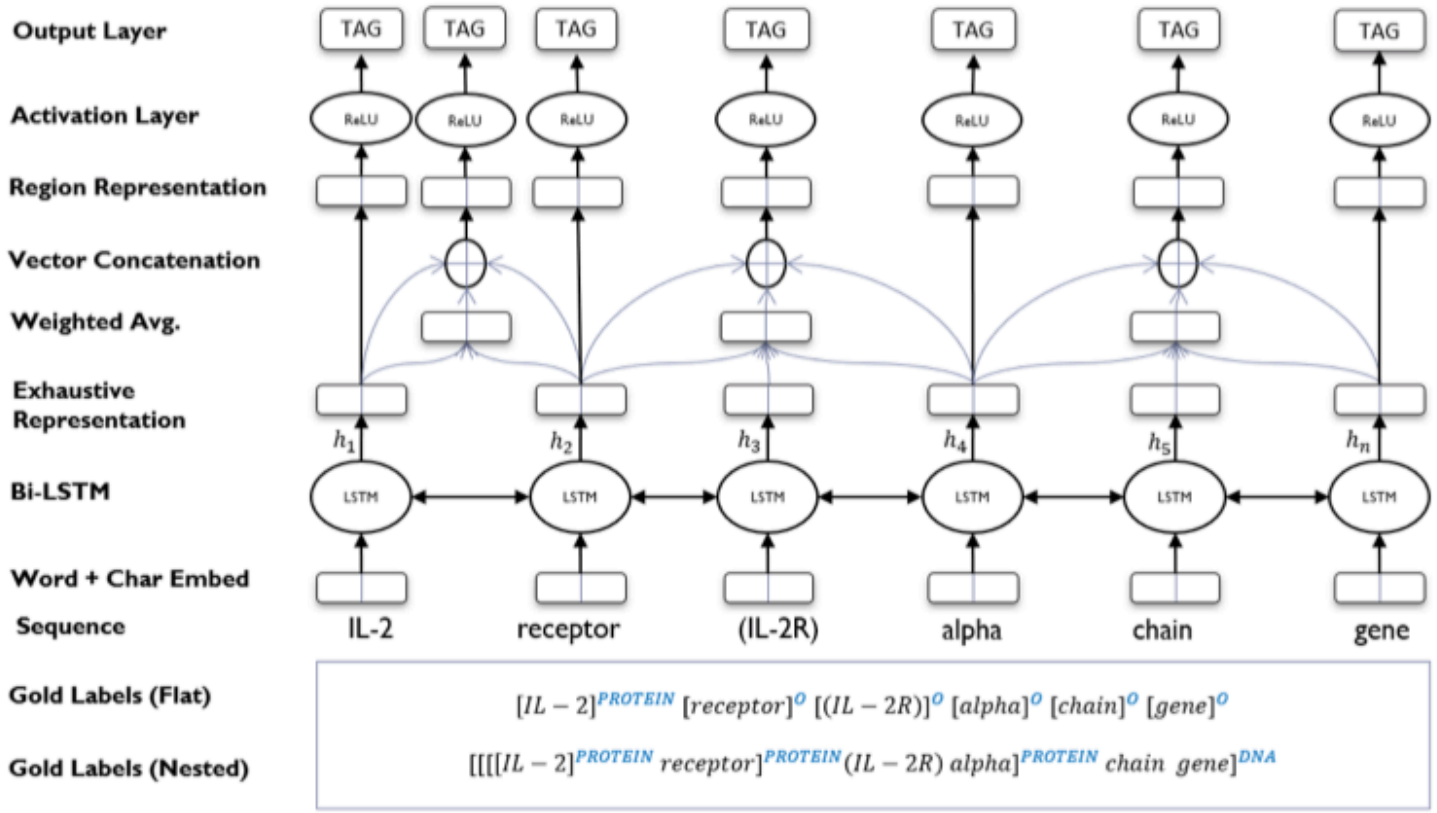
- Give a sequence $X = \{x_1, x_2, \dots\dots, x_n\}$
- Enumerate all sub-sequences of X,

$$S = \{s_1, s_1 s_2, s_1 s_2 s_3, \dots\dots, s_2, s_2 s_3, \dots\dots, s_n\}$$

- Train a Classifier C, predict the label of each sub-sequence:

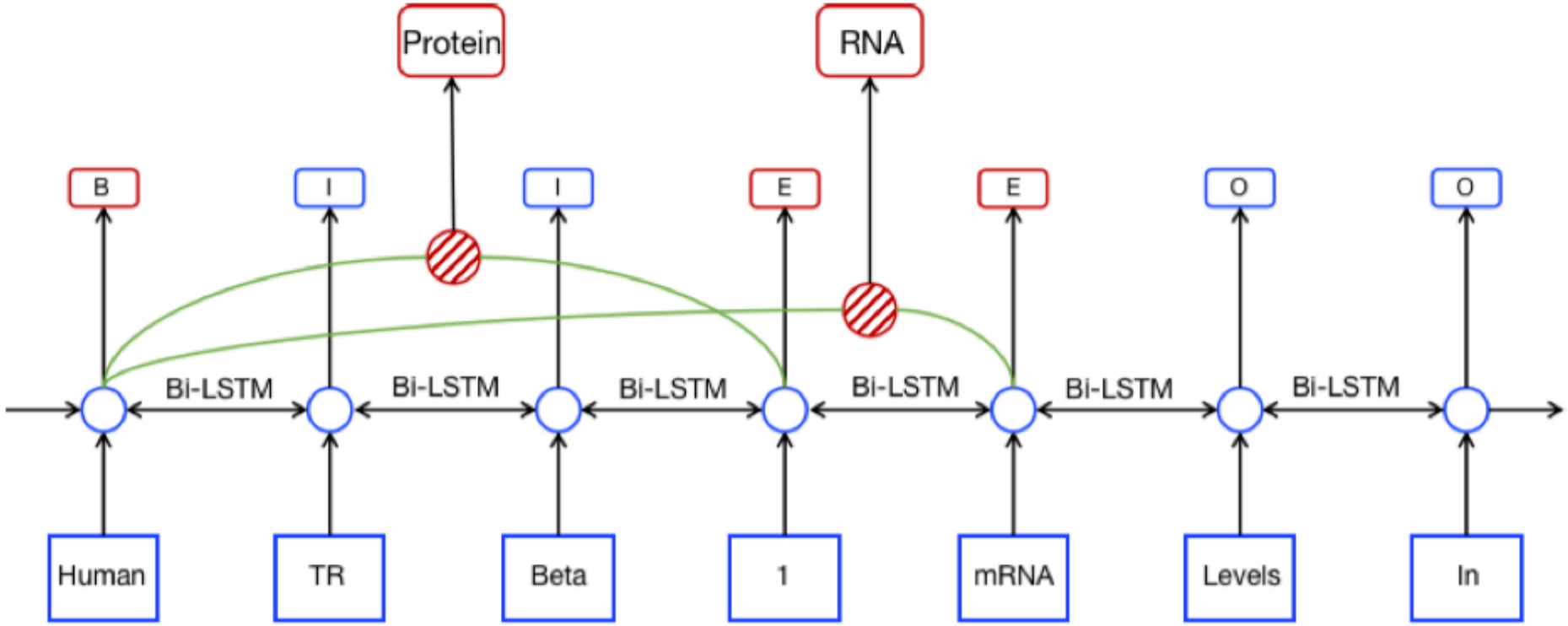
$$Y = \{y_1, y_2, \dots, y_m\}, \qquad m = len(S)$$

Deep Exhaustive Model



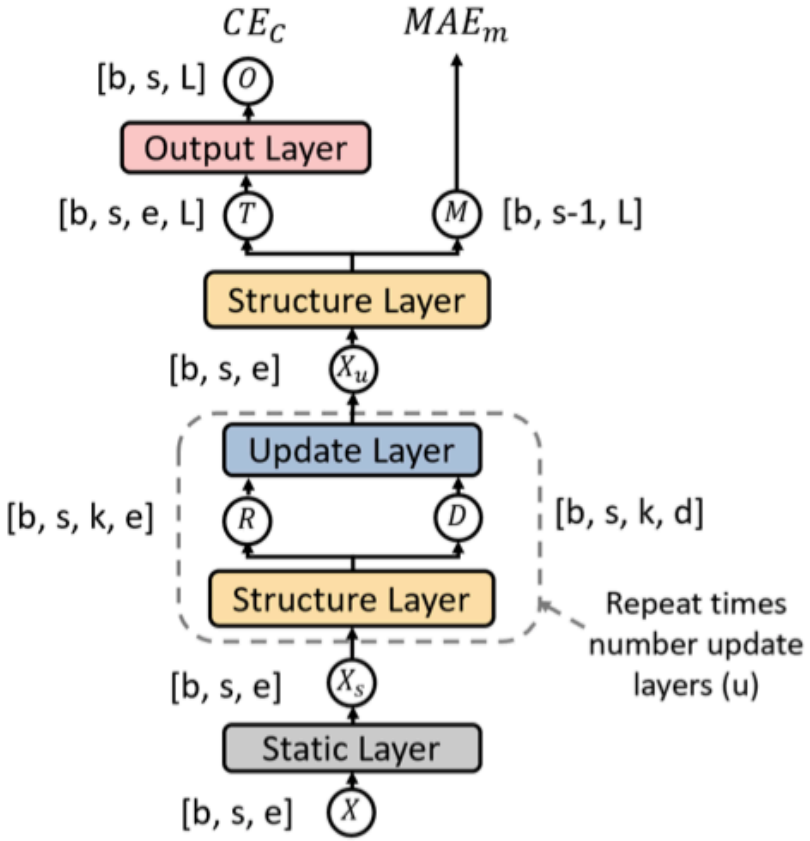
Deep Exhaustive Model for Nested Named Entity Recognition, EMNLP 2018

Boundary-aware Model



A Boundary-aware Neural Model for Nested Named Entity Recognition, EMNLP-IJCNLP 2019

Connection-aware Model



Merge and Label: A Novel Neural Network Architecture for Nested NER. ACL 2019

Problems

- Negative Samples
- Length of sub-sequence
- High complexity

Results on GENIA

	P	R	F
Neural Layered	78.5	71.3	74.7
Deep Exhaustive	73.3	68.3	70.7
Linearization(Flair)	/	/	78.31
Boundary-aware	75.9	73.6	74.7

Results on ACE2005

	P	R	F
Neural Layered	74.2	70.3	72.2
Linearization(Flair)	/	/	84.33
Merge and Label	82.7	82.1	82.4

Summary

- The trend is use sub-sequence classifier
- Most work aims to reduce the number of negative samples and complexity

- Layered model still work
- Combined label performs well due to BERT/ELMo.