Pre- and In-Parsing Models for Neural Empty Category Detection

Yufei Chen, Yuanyuan Zhao, Weiwei Sun and Xiaojun Wan

Institute of Computer Science and Technology
Peking University

Deep Linguistic Processing

- Combinatory Categorial Grammar, Lexical-Functional Grammar, Head-Driven Phrase-Structure Grammar, etc.
- Government and Binding: D-structure VS S-structure

Question: Can we benefit from moding deep elements?

- Perhaps. Deep grammar formalisms provide more transparent interface to semantics
- Hard to prove. Grammar Formalism are heterogeneous and hard to be compared.
- Modeling empty category help dependency parsing.
  - Our CoNLL paper: Zhang, Sun and Wan (2017)
  - The dependency tree representation is augmented with *empty nodes*, which corresponds to unpronounced nominal *words*
  - Data-driven parsing based on global linear models

Question: How about neural models?

- Is is plausible to detect empty categories using RNNs rather than syntactic information?
- Can neural parsing benefit from modeling empty categories?

Pre-parsing neural empty category detection

- Context of empty categories: sequential context and hierarchical context
- A sequence-oriented model: we explore four sets of annotation specifications
- Tagging based on a BiLSTM-CRF model.

Interspace:	@@ 颁布(issue) O VV	@@ 了(AS) O AS	@@ 涉及(involve) *OP**T* VV	@@ 经济(economic) O NN
Pre2 and *Pre3*:	颁布(issue) VV	了(AS) AS	涉及(involve) VV#pre1=*T*#pre2=*OP*	经济(economic) NN
Prepost:	颁布(issue) VV	了(AS) AS#post=*OP*	涉及(involve) VV#pre1=*T*	经济(economic) NN

Figure 1: An example of four kinds of annotations. "@@" means interspaces between words.

Joint ECD and dependency parsing

- Notation
  - a sentence s with n normal words
  - $\mathcal{I}_o = \{(i,j)|i,j \in \{1,\cdots,n\}\}$: all possible overt dependency edges
  - $\mathcal{I}_c = \{(i,\phi_j)|i,j \in \{1,\cdots,n\}\}$: all possible covert dependency edges. ϕ_j denotes an empty node that precede the jth word.
  - $z = \{z(i,j) : (i,j) \in \mathcal{I}_o \cup \mathcal{I}_c\}$: a dependency parse with empty nodes
- Parsing with ECD can be defined as a search for the highest-scored $z^*(s)$ in all compatible analyses, just like parsing without empty elements:

$$z^*(s) = \arg\max_{z \in \mathcal{Z}(s)} \textsc{Score}(s,z)$$
$$= \arg\max_{z \in \mathcal{Z}(s)} \sum_{p \in \textsc{Part}(z)} \textsc{ScorePart}(s,p)$$

A second-order model

the score function over the whole syntactic analysis is defined as:

$$\begin{aligned}\textsc{Score}(s,\boldsymbol{z}) = &\sum_{(i,j)\in\textsc{Dep}(\boldsymbol{z})} \textsc{ScoreDep}(s,i,j) \\ &+ \sum_{(i,\phi_j)\in\textsc{DepEmpty}(\boldsymbol{z})} \textsc{ScoreEmpty}(s,i,\phi_j) \\ &+ \sum_{(i,j,k)\in\textsc{OvertBoth}(\boldsymbol{z})} \textsc{ScoreOvertBoth}(s,i,j,k) \\ &+ \sum_{(i,\phi_j,k)\in\textsc{CovertIn}(\boldsymbol{z})} \textsc{ScoreCovertIn}(s,i,\phi_j,k) \\ &+ \sum_{(i,j,\phi_k)\in\textsc{CovertOut}(\boldsymbol{z})} \textsc{ScoreCovertOut}(s,i,j,\phi_k)\end{aligned}$$

The score functions

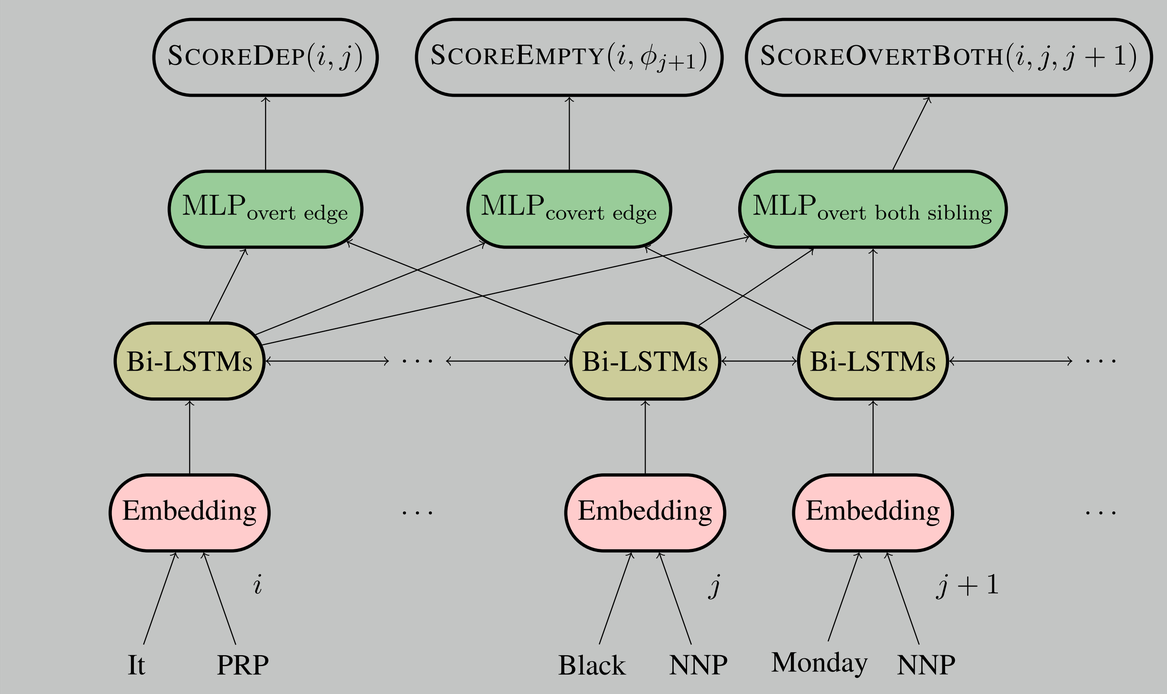

Figure 2: The neural network structure when parsing sentence "It wasn't Black Monday." 5 MLPs is used for overt edges (i,j), covert edges (i,ϕ_j), overt-both siblings (i,j,k), covert-inside siblings (i,ϕ_j,k) and covert-outside siblings (i,j,ϕ_k) respectively, and 3 of them are shown in the graph.

Overall results

	P	R	F_1
Pre-parsing	67.3	54.7	60.4
In-parsing	72.6	55.5	62.9
In-parsing*	70.9	54.1	61.4
Xue and Yang (2013)*	65.3	51.2	57.4
Cai et al. (2011)	66.0	54.5	58.6

Table 1: The overall performance on test data. "*" indicates more stringent evaluation metrics.

Empty category helps neural parsing

	−EC	+EC	−+EC
Unlabeled	87.6	88.9	89.6
Labeled	84.6	85.9	86.6

Table 2: Accuracies of both unlabeled and labeled parsing on development data. "−EC" indicates parsing without empty categories. "+EC" indicates the second-order in-parsing models. "−+EC" indicates jointing parsing models both without and with ECs together.

LSTM is able to find some non-local dependencies

	Linear CRF						LSTM-CRF					
	Without POS			With POS			Without POS			With POS		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
Interspace	74.6	20.6	32.2	71.2	30.3	42.5	67.9	59.8	63.6	73.0	61.6	66.8
Pre2	72.4	30.1	42.5	72.8	32.4	44.8	71.1	58.3	64.1	74.8	57.4	65.0
Pre3	73.1	30.2	42.8	73.0	32.5	44.9	71.1	58.5	64.2	73.8	57.0	64.3
Prepost	70.9	32.9	45.0	74.4	30.3	43.1	71.0	57.6	63.6	72.9	58.6	65.0

Table 3: The overall performance of the two sequential models on development data.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (61772036, 61331011) and the Key Laboratory of Science, Technology and Standard in Press Industry (Key Laboratory of Intelligent Press Media Technology).

Contact Information

Email: {yufei.chen,ws,wanxiaojun}@pku.edu.cn