Dependency Grammar Induction with Neural Lexicalization and Big Training Data

Wenjuan Han · Yong Jiang · Kewei Tu

School of Information Science and Technology, ShanghaiTech University

Background

Grammar induction can benefit from incorporating lexical information into the learned grammar. In order to mitigate the data scarcity problem of full lexicalization, partial lexicalization (Headden et al., 2009), large data corpora (Pate and Johnson, 2016) and smoothing techniques (Jiang et al., 2016) are proposed in the literature.

In this work:

- We study the impact of big models (in terms of the degree of lexicalization) and big training data (in terms of the training corpus size) on the accuracy of grammar induction approaches. We experimented with two models: L-DMV, a lexicalized version of Dependency Model with Valence (Klein and Manning, 2004) and L-NDMV, our lexicalized extension of the Neural Dependency Model with Valence (Jiang et al., 2016).
- When trained on our biggest corpus, L-NDMV with a moderate degree of lexicalization and good model initialization achieves competitive performance with the current state-of-the-art.

Methods

L-DMV:

DMV (the Dependency Model with Valence (Klein and Manning, 2004):

- Generative model of sentences.
- Three types of rules: `CHILD`, `DECISION` and `ROOT`.
- The objective function is the negative log-likelihood of the training sentences $X = \{x_1, x_2, ..., x_n\}$.

$$L = -\sum_{\alpha=1}^{n} \log \sum_{y_\alpha \in \mathcal{Y}} \mathrm{P}(x_\alpha, y_\alpha)$$

$$\mathrm{P}(x_\alpha, y_\alpha) = \prod_{r \in R_\alpha(x_\alpha, y_\alpha)} p(r)$$

$\mathrm{P}(x_\alpha, y_\alpha)$ is the multiplication of all the rules that can be used when generating (x_α, y_α). $\mathcal{Y}$ is the set of all valid trees.

- Usually unlexicalized and based on POS tags.

L-DMV :

- Using partial lexicalization in which each token is represented by a word-POS pair (an infrequent word is represented by its POS tag).

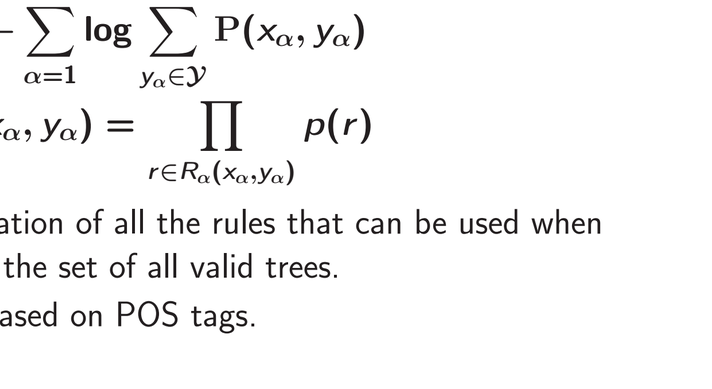
L-NDMV:

Difference from L-DMV:

- The probability of rule r is obtained by a neural network. The parameters of the neural network are denoted by Θ.

Two outputs of the neural network given the head word, head POS tag, direction and valence:

- Probabilities of `CHILD` rules $[p_{c_1}, p_{c_2}, ..., p_{c_m}]$ (m is the vocabulary size; c_i is the i-th token).
- Probabilities of `DECISION` rules $[p_{stop}, p_{continue}]$.



Model Initialization:

We tested two kinds of initialization:

1. KM initialization (Klein and Manning, 2004).
2. 'Expert' initialization: First learn an unlexicalized DMV using the grammar induction method of Naseem et al. (2010) and use it to parse the training corpus; then, from the parse trees we run maximum likelihood estimation to produce the initial lexicalized model.

Study 1 : Big Training Data—Increase Sentences Number from About 5k to 180k

Training corpora (sentences with length less than 10):

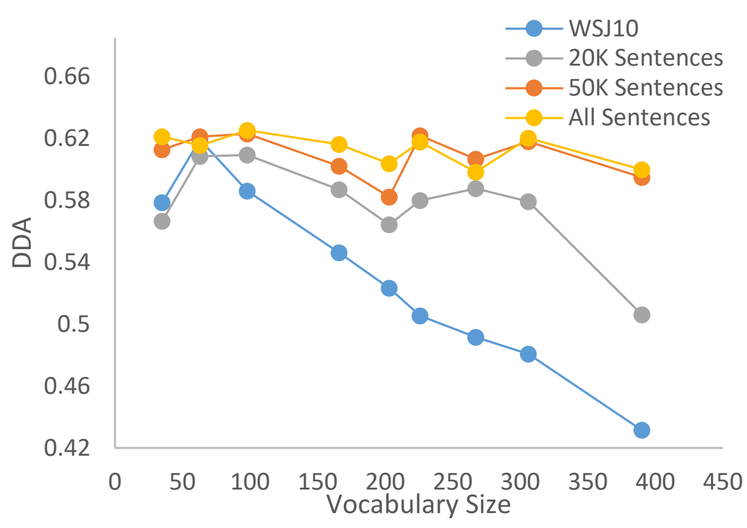
- WSJ corpus (about 5k).
- 20k, 50k and 180k sentences picked randomly from the BLLIP corpus.

Horizontal axis:

- The degree of lexicalization (here we control the degree by replacing words that appear less than a cutoff number with their POS tags. Different cut off numbers result in different vocabulary sizes).

Vertical axis:

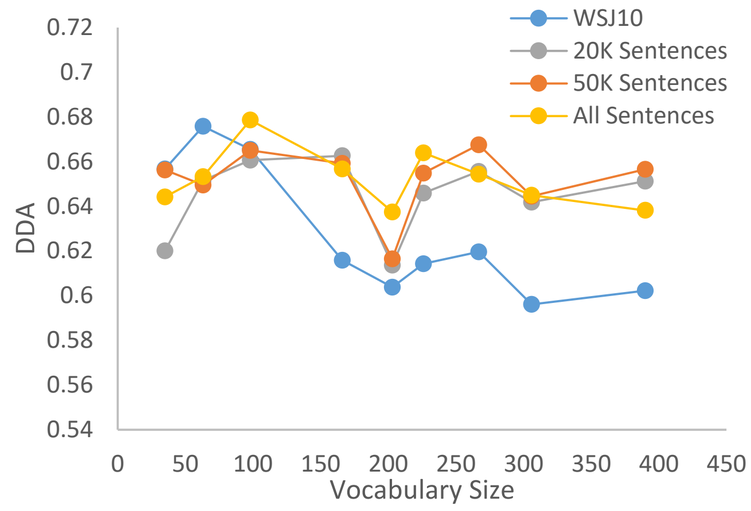
- The directed dependency accuracy (DDA).



Method: L-DMV with KM initialization:

- Severe degradation can be observed with more lexicalization.
- No benefit from the largest training corpus.

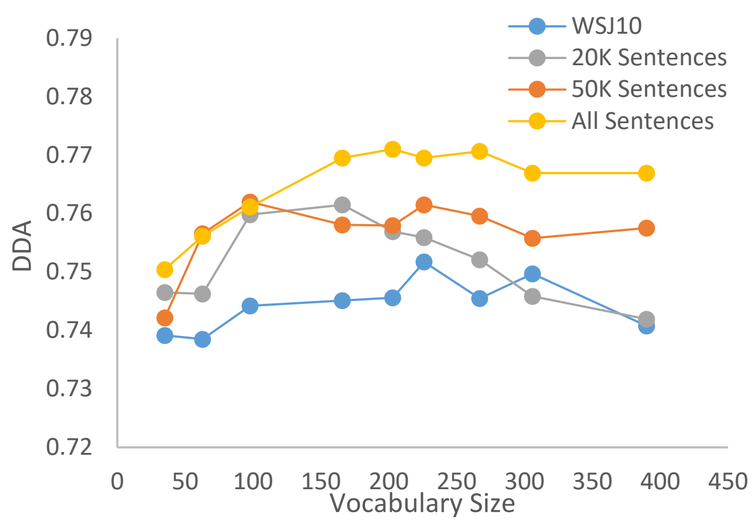
Study 2 : Neural Lexicalization—Use Neural Networks for Smoothing



Method: L-NDMV with KM initialization:

- Degradation with large degrees of lexicalization is much less severe.
- Diminishing return of big data can still be observed.

Study 3 : Use 'Expert' Initialization



Method: L-NDMV with 'expert' initialization:

- Accuracy does not decrease until the highest degrees of lexicalization.
- Better accuracy with increasing data sizes.

Comparison with Previous Systems

L-NDMV (with the largest corpus and the vocabulary size of 203 which was selected on the validation set) is competitive with previous state-of-the-art approaches.

Methods	WSJ10	WSJ
Unlexicalized Approaches, with WSJ10		
Neural E-DMV (Jiang et al., 2016)	72.5	57.6
Systems Using Lexical Information and/or More Data		
LexTSG-DMV (Blunsom and Cohn, 2010)	67.7	55.7
L-EVG (Headden III et al., 2009)	68.8	-
CS (Spitkovsky et al., 2013)	72.0	64.4
MaxEnc (Le and Zuidema, 2015)	73.2	**65.8**
L-NDMV + WSJ	75.1	59.5
L-NDMV + Large Corpus	**77.2**	63.2