# Combining Generative and Discriminative Approaches to Unsupervised Dependency Parsing via Dual Decomposition

Yong Jiang · Wenjuan Han · Kewei Tu
School of Information Science and Technology, ShanghaiTech University

## Motivation

The generative LC-DMV model and the discriminative Convex-MST model have achieved SOTA performance on unsupervised dependency parsing. So We study joint training of the two models to combine their strength.

## Proposed Solution

We proposed a decoding based training procedure via dual decomposition to effectively enjoy the benefits of the two models.

## Empirical Results

A State-of-The-Art performance on the UD 1.4 dataset is achieved over thirty languages.

## Unsupervised Dependency Parsing

Dependency parsing pipeline:

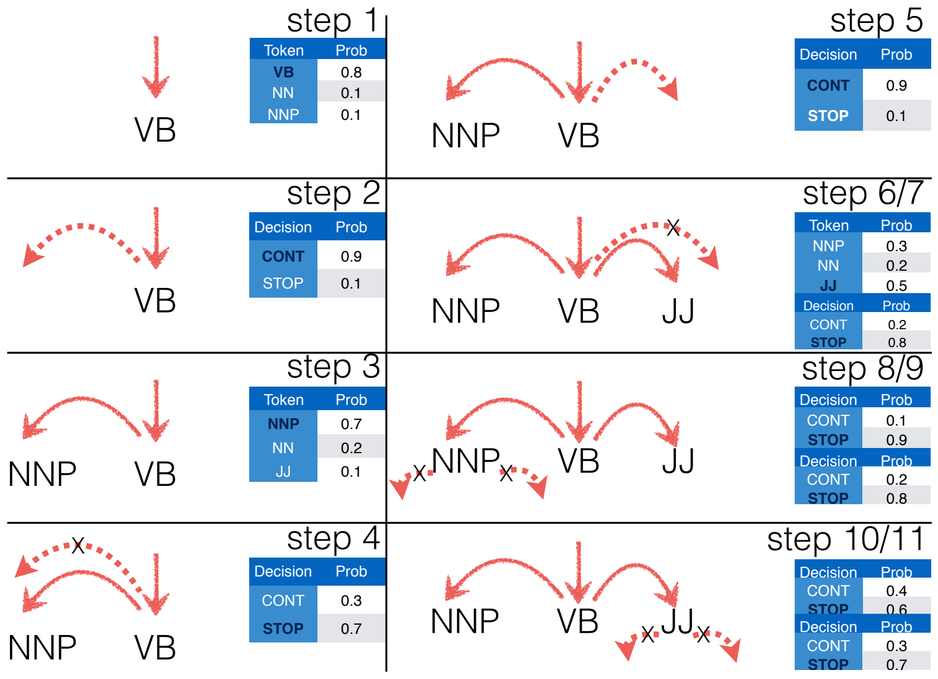

Unsupervised dependency parsing:

- No access to golden trees.
- Usually taking POS tags as inputs.
- Typically incorporating different types of inductive biases.

## LC-DMV

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

- A generative model of sentences in a top down manner.
- Three kinds of grammar rules: `CHILD`, `DECISION` and `ROOT`.
- An example is shown in below:




Learning of the DMV model:

- **Joint probability:** $P(\mathbf{x}, \mathbf{y}) = \prod_{r \in \mathcal{R}(\mathbf{x},\mathbf{y})} p(r)$
- **Marginal probability:** $P(\mathbf{x}) = \sum_{y \in \mathcal{Y}(\mathbf{x})} \prod_{r \in \mathcal{R}(\mathbf{x},\mathbf{y})} p(r)$
- Training: EM algorithm.

The Left Corner Dependency Model with Valence (LC-DMV) (Noji et al., 2016)

- adding a constraint that limits center embedding to the model.
- encouraging short dependencies.

## Convex-MST

Convex-MST (Grave and Elhadad, 2015) is a discriminative model for unsupervised dependency parsing based on the first-order maximum spanning tree dependency parser (McDonald et al., 2005).

Representation:

- score of an edge: $\mathbf{w}^T \mathbf{f}(\mathbf{x}, i, j)$
- $\boldsymbol{f}$ is a hand-crafted feature template.

Learning:

- Learning is based on discriminative clustering:
- $\min_{\mathbf{y}_1, \mathbf{y}_2 \ldots, \mathbf{y}_N} \min_{\mathbf{w}} \frac{1}{N} \sum_{\alpha=1}^{N} \left( \frac{1}{2n_\alpha} ||\mathbf{y}_\alpha - \mathbf{X}_\alpha \mathbf{w}||_2^2 - \mu \mathbf{v}^T \mathbf{y}_\alpha \right) + \frac{\lambda}{2} ||\mathbf{w}||_2^2$

Decoding:

- Intractable!
- Approximated by finding a continuous solution and then round to a discrete solution.
  - check out our EMNLP 2017 paper on a tractable CRF-Autoencoder model (Cai et al., 2017)!

## Jointly Training with Dual Decomposition

Pros and Cons of the Generative Model:

- Pros: Learning is quite easy. Many inductive biases can be employed in the model.
- Cons: Global features are hard to encode in the model.

Pros and Cons of the Discriminative Model:

- Pros: Enjoying global features.
- Cons: Some inductive bias (e.g: center embedding) is hard to encode in the model.

Jointly optimize the two models with a combined objective function:

$$J(\mathbf{M_F}, \mathbf{M_G}) = \sum_{\alpha=1}^{N} \min_{\mathbf{y}_\alpha \in \mathcal{Y}_\alpha} \left( F(\mathbf{x}_\alpha, \mathbf{y}_\alpha; \mathbf{M_F}) + G(\mathbf{x}_\alpha, \mathbf{y}_\alpha; \mathbf{M_G}) \right)$$

where the two components are:

$$F(\mathbf{x}_\alpha, \mathbf{y}_\alpha; \Theta) = -\log \left( P_\Theta(\mathbf{x}_\alpha, \mathbf{y}_\alpha) f(\mathbf{x}_\alpha, \mathbf{y}_\alpha) \right)$$

$$G(\mathbf{x}_\alpha, \mathbf{y}_\alpha; \mathbf{w}) = \frac{1}{2n_\alpha} ||\mathbf{y}_\alpha - \mathbf{X}_\alpha \mathbf{w}||_2^2 + \frac{\lambda}{2N} ||\mathbf{w}||_2^2 - \mu \mathbf{v}^T \mathbf{y}$$

Learning via Coordinate Descent:

**Algorithm 1** Parameter Learning

**Input:** Training sentence $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N$
Pre-train $\Theta$ and $\mathbf{w}$
**repeat**
  Fix $\Theta$ and $\mathbf{w}$ and solve the decoding problem to get $\mathbf{y}_\alpha, \alpha = 1, 2, \ldots, N$
  Fix the parses and update $\Theta$ and $\mathbf{w}$
**until** Convergence

**Algorithm 2** Decoding via Dual Decomposition

**Input:** Sentence $\mathbf{x}$, fixed parameters $\mathbf{w}$ and $\Theta$
Initialize vector $\mathbf{u}$ of size $n \times n$ to $\mathbf{0}$
**repeat**
  $\hat{\mathbf{y}} = \arg\min_{\mathbf{y} \in \mathcal{Y}} F(\mathbf{x}, \mathbf{y}; \Theta) + \mathbf{u}^T \mathbf{y}$
  $\hat{\mathbf{z}} = \arg\min_{\mathbf{z} \in \mathcal{Y}} G(\mathbf{x}, \mathbf{z}; \mathbf{w}) - \mathbf{u}^T \mathbf{z}$
  **if** $\hat{\mathbf{y}} = \hat{\mathbf{z}}$ **then**
    return $\hat{\mathbf{y}}$
  **else**
    $\mathbf{u} = \mathbf{u} - \tau (\hat{\mathbf{y}} - \hat{\mathbf{z}})$
  **end if**
**until** Convergence

## Experiments and Analysis

**Dataset:** Thirty languages on Universal Dependencies (UD) Treebank 1.4.

- Training data: Sentence length no more than 15.
- Testing data: Sentence length no more than 45.

| Language | M | D | D-I | M-J | D-J | DD |
|---|---|---|---|---|---|---|
| A_Greek | 43.4 | 33.1 | 38.8 | 44.2 | **44.9** | 38.9 |
| A_Greek-P | 50.4 | 43.0 | 44.7 | 50.8 | **52.9** | 44.9 |
| Basque | 50.0 | 45.4 | 54.2 | 52.1 | **55.7** | 50.2 |
| Bulgarian | 61.6 | 62.4 | 60.3 | 64.7 | **73.8** | 64.8 |
| Czech | 48.6 | 17.4 | 53.9 | 48.7 | **54.0** | 53.5 |
| Czech-CAC | 50.4 | 53.0 | 53.9 | 55.6 | **62.3** | 50.2 |
| Dutch | 45.3 | 34.1 | **56.7** | 48.2 | 43.5 | 40.7 |
| Dutch-LS | 42.4 | 27.0 | 16.4 | **43.2** | 41.2 | 36.3 |
| English | 54.0 | 56.0 | 49.8 | 57.3 | **60.1** | 53.4 |
| Estonian | **49.4** | 31.8 | 47.5 | 48.7 | 44.0 | 44.4 |
| Finnish | **44.7** | 26.9 | 39.0 | 44.2 | 43.5 | 31.2 |
| Finnish-FTB | **49.9** | 31.0 | 47.9 | 47.7 | 48.0 | 36.5 |
| French | **62.0** | 48.6 | 57.0 | 54.5 | 57.0 | 55.5 |
| German | 51.4 | 50.5 | 54.1 | 49.3 | **55.7** | 48.6 |
| Gothic | 52.7 | 49.9 | 47.3 | **59.6** | 56.4 | 58.0 |
| Hindi | 56.8 | 54.2 | 48.4 | 52.1 | **60.0** | 49.1 |
| Italian | 69.1 | **71.1** | 67.4 | 62.8 | 70.3 | 64.5 |
| Japanese | 44.8 | 43.8 | 43.8 | 42.8 | **45.8** | 41.0 |
| Latin-ITTB | 38.8 | 38.6 | 42.3 | **47.0** | 42.2 | 40.3 |
| Latin-PROIEL | 44.3 | 34.8 | 38.7 | **46.8** | 41.8 | 42.9 |
| Norwegian | 55.3 | 45.5 | 51.4 | 57.4 | **60.8** | 46.6 |
| Old_Church_S | 56.4 | 26.6 | 51.3 | 58.3 | **58.6** | 42.0 |
| Polish | 63.4 | 63.7 | 61.5 | 70.7 | **74.2** | 68.9 |
| Portuguese | 57.9 | **67.2** | 60.1 | 56.1 | 62.9 | 57.4 |
| Portuguese-BR | 59.3 | 63.1 | 62.0 | 65.5 | **68.8** | 58.3 |
| Russian-STR | 47.6 | 51.7 | 56.5 | 52.1 | **64.4** | 52.6 |
| Slovak | 57.4 | 59.3 | 51.9 | 61.7 | **65.9** | 58.7 |
| Slovenian | 54.0 | 49.5 | 56.3 | 65.5 | **69.6** | 56.1 |
| Spanish | 61.9 | 61.9 | 60.3 | 57.4 | **68.0** | 60.2 |
| Spanish-AC | 59.4 | 59.5 | 56.4 | 56.8 | **65.2** | 57.6 |
| Average | 52.7 | 47.2 | 50.3 | 54.2 | **56.5** | 49.6 |
| Average $\leq 15$ | 55.4 | 48.9 | 54.9 | 57.3 | **60.2** | 53.8 |

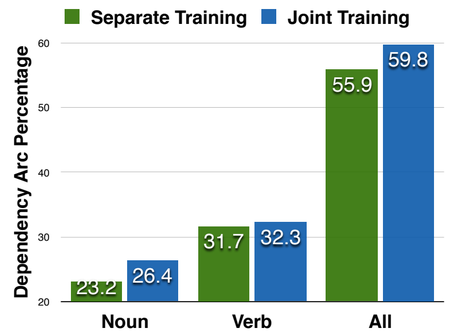



Figure 1: Percentages of dependencies satisfying linguistic rules in the LC-DMV parses of the English test dataset. Noun and Verb denote dependencies headed by nouns and verbs.

| Methods | Average Dependency Length |
|---|---|
| Separate Training | 1.673 |
| Joint Training | **1.627** |

Table 2: Average dependency length in the Convex-MST parses of the English test dataset.

In the left table:

- **M:** Convex-MST; **D:** LC-DMV
- **D-I:** LC-DMV initialized from results of **M**
- **M-J:** Convex-MST model trained jointly
- **D-J:** LC-DMV model trained jointly
- **DD:** Parses decoded jointly

## Future Work

- Adding more word information together with POS tags. (see our EMNLP 2017 paper on adding more lexical info (Han et al., 2017))
- Applying unsupervised models to semi-supervised parsing.