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 propose 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.

### 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.

![Image of dependency parsing](image)

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

<table>
<thead>
<tr>
<th>Dependency Type</th>
<th>Separate Training</th>
<th>Joint Training</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noun-Verb</td>
<td>10.1</td>
<td>6.2</td>
</tr>
<tr>
<td>Verb-Noun</td>
<td>3.0</td>
<td>1.7</td>
</tr>
<tr>
<td>Noun-Noun</td>
<td>37.5</td>
<td>34.2</td>
</tr>
<tr>
<td>Verb-Verb</td>
<td>20.3</td>
<td>16.9</td>
</tr>
<tr>
<td>Noun-Pronoun</td>
<td>9.2</td>
<td>8.7</td>
</tr>
<tr>
<td>Pronoun-Noun</td>
<td>6.2</td>
<td>5.9</td>
</tr>
<tr>
<td>Pronoun-Pronoun</td>
<td>13.2</td>
<td>12.4</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Dependency Type</th>
<th>Separate Training</th>
<th>Joint Training</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noun-Verb</td>
<td>10.1</td>
<td>6.2</td>
</tr>
<tr>
<td>Verb-Noun</td>
<td>3.0</td>
<td>1.7</td>
</tr>
<tr>
<td>Noun-Noun</td>
<td>37.5</td>
<td>34.2</td>
</tr>
<tr>
<td>Verb-Verb</td>
<td>20.3</td>
<td>16.9</td>
</tr>
<tr>
<td>Noun-Pronoun</td>
<td>9.2</td>
<td>8.7</td>
</tr>
<tr>
<td>Pronoun-Noun</td>
<td>6.2</td>
<td>5.9</td>
</tr>
<tr>
<td>Pronoun-Pronoun</td>
<td>13.2</td>
<td>12.4</td>
</tr>
</tbody>
</table>

**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.

---

**Motivation**

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 the following:

**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 the following:

**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).

**Learning:**
- Learning is based on discriminative clustering:
  \[
  \min_{w_0, w_1, \ldots, w_N} \frac{1}{N} \sum_{n=1}^{N} \left( \frac{1}{2 \Delta n} ||y_{n} - X_n w||^2 - \mu v^T y_n \right) + \frac{1}{2} ||w||^2
  \]
  \[
  \text{Decoding:}
  \text{Intractable!}
  \text{Approximated by finding a continuous solution and then round to a discrete solution.}
  \text{Check out our EMNLP 2017 paper on a tractable CRF-Autoencoder model (Cai et al., 2017).}

---

**Algorithm 1 Parameter Learning**

**Input:** Sentence \(x_1, x_2, \ldots, x_N\)

**Algorithm 2 Decoding via Dual Decomposition**

**Input:** Sentence \(x\), fixed parameters \(w\) and \(\Theta\)

**Learning via Coordinate Descent:**

- **Joint probability:** \(P(x, y) = \prod_{i \in \mathcal{R}(x, y)} P(r)\)
- **Marginal probability:** \(P(x) = \sum_{y \in \mathcal{Y}(x)} \prod_{i \in \mathcal{R}(x, y)} P(r)\)
- **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.