Second-Order Semantic Dependency Parsing with End-to-End Neural Networks

Xinyu Wang, Jingxian Huang, Kewei Tu*
School of Information Science and Technology, ShanghaiTech University, Shanghai, China
{wangxy1,huangjx,tukw}@shanghaitech.edu.cn

Abstract
Semantic dependency parsing aims to identify semantic relationships between words in a sentence that form a graph. In this paper, we propose a second-order semantic dependency parser, which takes into consideration not only individual dependency edges but also interactions between pairs of edges. We show that second-order parsing can be approximated using mean field (MF) variational inference or loopy belief propagation (LBP). We can unfold both algorithms as recurrent layers of a neural network and therefore can train the parser in an end-to-end manner. Our experiments show that our approach achieves state-of-the-art performance.

1 Introduction
Semantic dependency parsing (Oepen et al.) aims to produce graph-structured semantic dependency representations of sentences instead of tree-structured syntactic dependency parses. Existing approaches to semantic dependency parsing can be classified as graph-based approaches and transition-based approaches. In this paper, we investigate graph-based approaches which score each possible parse of a sentence by factorizing over its parts and search for the highest-scoring parse.

Previous work in graph-based syntactic dependency parsing has shown that higher-order parsing generally outperforms first-order parsing (McDonald and Pereira, 2006; Carreras, 2007; Koo and Collins, 2010; Ma and Zhao, 2012). While a first-order parser scores dependency edges independently, a higher-order parser takes relationships between two or more edges into consideration. However, most of the previous algorithms for higher-order syntactic dependency tree parsing are not applicable to semantic dependency graph parsing, and designing efficient algorithms for higher-order semantic dependency graph parsing is nontrivial. In addition, it becomes a common practice to use neural networks to compute features and scores of parse graph components, which ideally requires backpropagation of parsing errors through the higher-order parsing algorithm, adding to the difficulty of designing such an algorithm.

In this paper, we propose a novel graph-based second-order semantic dependency parser. Given an input sentence, we use a neural network to compute scores for both first and second-order parts of parse graphs and then apply either mean field variational inference or loopy belief propagation to approximately find the highest-scoring parse graph. Both algorithms are iterative inference algorithms and we show that they can be unfolded as recurrent layers of a neural network with each layer representing the computation in one iteration of the algorithms. In this way, we can construct an end-to-end neural network that takes in a sentence and outputs the approximate marginal probability of every possible dependency edge. During training, we maximize the probability of the gold parses by using standard gradient-based methods. Our experiments show that our approach achieves state-of-the-art performance in semantic dependency parsing and outperforms our baseline with 0.3% and 0.4% labeled F1 score and previous state-of-the-art model with 1.3% and 1.4% labeled F1 score for in-domain and out-of-domain test sets respectively. Our approach shows more advantage over the baseline when there are fewer training data and when parsing longer sentences.

2 Semantic Dependency Parsing

Broad-Coverage Semantic Dependency Parsing was first defined in SemEval-2014 task 8 (Oepen
et al.) aiming at recovering semantic dependency relationships in sentences of the WSJ corpus. It was extended in SemEval-2015 task 18 (Oepen et al., 2015) with an additional out-of-domain dataset (the Brown corpus). A semantic dependency parse is different from a syntactic dependency parse in that the dependency edges are annotated with semantic relations (e.g., agent and patient) and form a directed acyclic graph instead of a tree. The Broad-Coverage Semantic Dependency Parsing provides three different formalisms: DM, PAS and PSD. Previous work has found that PAS is the easiest to learn and PSD is the most difficult as it has the largest set of labels.

3 Approach

Our model architecture (shown in Figure 1) follows that of Dozat and Manning (2018). Given an input sentence, we first compute word representations using a BiLSTM, which are then fed into two parallel modules, one for predicting the existence of every edge and the other for predicting the label of every edge. The label-prediction module makes predictions of each edge independently and hence is a first-order decoder. The edge-prediction module is what our approach differs from that of Dozat and Manning (2018). The module scores both first and second-order parts and then goes through multiple recurrent inference layers to predict edge existence.

3.1 Part Scoring

Given a sentence with \( n \) words \( [w_1, w_2, ..., w_n] \), we feed a BiLSTM with their word embeddings and POS tag embeddings.

\[
o_i = e_i^{\text{word}} \oplus e_i^{\text{postag}}
R = \text{BiLSTM}(O)
\]

where \( o_i \) is the concatenation (\( \oplus \)) of the word and POS tag embeddings of word \( w_i \), \( O \) represents \( [o_1, \ldots, o_n] \), and \( R = [r_1, \ldots, r_n] \) represents the output from the BiLSTM.

To score first-order parts (edges) in both the edge-prediction module and the label-prediction module, we use two single-layer feedforward networks (FNNs) to compute a head representation and a dependent representation for each word and then apply a biaffine function to compute the scores of edges and labels.

\[
\text{Biaff}(v_1, v_2) := v_1^T U v_2 + b
\]

\[
h_i^{\text{edge-head}} = \text{FNN}^{\text{edge-head}}(r_i)
\]

\[
h_i^{\text{edge-dep}} = \text{FNN}^{\text{edge-dep}}(r_i)
\]

\[
h_i^{\text{label-head}} = \text{FNN}^{\text{label-head}}(r_i)
\]

\[
h_i^{\text{label-dep}} = \text{FNN}^{\text{label-dep}}(r_i)
\]

\[
s_{ij}^{\text{edge}} = \text{Biaff}^{\text{edge}}(h_i^{\text{edge-dep}}, h_j^{\text{edge-head}})
\]

\[
s_{ij}^{\text{label}} = \text{Biaff}^{\text{label}}(h_i^{\text{label-dep}}, h_j^{\text{label-head}})
\]

In Eq. 2, the tensor \( U \) in the biaffine function is \((d \times c \times d)\)-dimensional and is diagonal (for any
where $d$ is hidden size and $c$ is the number of labels. In Eq. 1, the tensor $U$ in the biaffine function is $d \times 1 \times d$-dimensional.

In the edge-prediction module, we further score second-order parts. We consider three types of second-order parts: siblings (sib), co-parents (cop) and grandparents (gp) (Martins and Almeida, 2014), as shown in Figure 2. For a specific type of second-order part, we use single-layer FNNs to compute a head representation and a dependent representation for each word. For grandparent parts, we additionally compute a head_dep representation for each word.

\[
type \in \{ \text{sib, cop, gp} \}
\]

\[
h^\text{(type-head)}_i = \text{FNN}(\text{type-head})(r_i)
\]

\[
h^\text{(type-dep)}_i = \text{FNN}(\text{type-dep})(r_i)
\]

\[
h^\text{(gp-head_dep)}_i = \text{FNN}(\text{gp-head_dep})(r_i)
\]

We then apply a trilinear function to compute scores of second-order parts. A trilinear function is defined as follows.

\[
\text{Trilin}(v_1, v_2, v_3) := v_3^T v_1^T U v_2
\]

where $U$ is a $(d \times d \times d)$-dimensional tensor. To reduce the computation cost, we assume that $U$ has rank $d$ and can be represented as the product of three $(d \times d)$-dimensional matrices $U_1$, $U_2$ and $U_3$. We can then compute second-order part scores as follows.

\[
g_i := U_i v_i \quad i \in [1, 2, 3]
\]

\[
\text{Trilin}(v_1, v_2, v_3) := \sum_{i=1}^{d} g_{1i} \circ g_{2i} \circ g_{3i}
\]

\[
s^\text{(sib)}_{ij,ik} = s^\text{(sib)}_{ik,ij} = \text{Trilin}(\text{h}^\text{(head)}_i, \text{h}^\text{(dep)}_j, \text{h}^\text{(dep)}_k)
\]

\[
s^\text{(cop)}_{ij,kj} = s^\text{(cop)}_{kj,ij} = \text{Trilin}(\text{h}^\text{(head)}_i, \text{h}^\text{(dep)}_j, \text{h}^\text{(head)}_k)
\]

\[
s^\text{(gp)}_{ij,jk} = \text{Trilin}(\text{h}^\text{(head)}_i, \text{h}^\text{(head-dep)}_j, \text{h}^\text{(dep)}_k)
\]

where $\circ$ represents element-wise product. We require $j < k$ in Eq. 4 and $i < k$ in Eq. 5.

\[
\phi_u(X_{ij}) = \begin{cases} 
\exp(s^\text{(edge)}_{ij}) & X_{ij} = 1 \\
1 & X_{ij} = 0
\end{cases}
\]

For each pair of edges $(i,j)$ and $(k,l)$ that form a second-order part of a specific type, we define a binary potential $\phi_p(X_{ij}, X_{kl})$.

\[
\phi_p(X_{ij}, X_{kl}) = \begin{cases} 
\exp(s^\text{(type)}_{ij,kl}) & X_{ij} = X_{kl} = 1 \\
1 & \text{Otherwise}
\end{cases}
\]

Exact inference on this CRF is intractable. We resort to iterative approximate inference algorithms as described below, which produce the posterior distribution $Q_{ij}(X_{ij})$ of each edge $(i,j)$. We can then predict the parse graph by including
every edge \((i, j)\) such that \(Q_{ij}(1) > 0.5\). The edge labels are predicted by maximizing the label probabilities computed by the label-prediction module.

**Mean Field Variational Inference**

Mean field variational inference approximates a true posterior distribution with a factorized variational distribution and tries to iteratively minimize their KL divergence. We can derive the following iterative update equations of distribution \(Q_{ij}(X_{ij})\).

\[
F_{ij}^{(t-1)} = \sum_{k \neq i,j} Q_{ik}^{(t-1)}(1) s_{ij,ik}^{( UIG )} + Q_{kj}^{(t-1)}(1) s_{ij,kj}^{( COP )}
\]

\[
+ Q_{jk}^{(t-1)}(1) s_{ij,jk}^{( UIG )} + Q_{ki}^{(t-1)}(1) s_{ij,ki}^{( COP )}
\]

\[
Q_{ij}^{(t)}(0) \propto 1
\]

\[
Q_{ij}^{(t)}(1) \propto \exp \{ s_{ij}^{(edge)} + F_{ij}^{(t-1)} \}
\]

The initial distribution \(Q_{ij}^{(0)}(X_{ij})\) is set by normalizing the unary potential \(\phi_u(X_{ij})\). We iteratively update the distributions for \(T\) steps and then output \(Q_{ij}^{(T)}(X_{ij})\), where \(T\) is a hyperparameter.

**Loopy Belief Propagation**

Loopy belief propagation iteratively passes messages between variables and potential functions (factors). Because our CRF contains only unary and binary potentials, we can merge each variable-to-factor message and its subsequent factor-to-variable message \(M_{kl \rightarrow ij}\), representing message from edge \((k, l)\) to edge \((i, j)\). The update function of the messages in each iteration is:

\[
Q_{ij}^{(t)}(X_{ij}) = \phi_u(X_{ij}) \prod_{ab \in \mathcal{X}_{ij}} M_{ab \rightarrow ij}^{(t-1)}(X_{ij})
\]

\[
M_{kl \rightarrow ij}^{(t)}(0) \propto \sum_{x_{kl}} Q_{kl}^{(t-1)}(x_{kl})/M_{ij \rightarrow kl}^{(t-1)}(x_{kl})
\]

\[
M_{kl \rightarrow ij}^{(t)}(1) \propto Q_{kl}^{(t-1)}(0)/M_{ij \rightarrow kl}^{(t-1)}(0)
\]

\[
+ \exp(s_{ij,kl}^{(type)}) Q_{kl}^{(t-1)}(1)/M_{ij \rightarrow kl}^{(t-1)}(1)
\]

We initialize the messages with \(M_{kl \rightarrow ij}^{(0)} = 1\). We iteratively update the messages and distributions for \(T\) steps and then output \(Q_{ij}^{(T)}(X_{ij})\).

**Inference as Recurrent Layers**

Zheng et al. (2015) proposed that a fixed number of iterations in mean field variational inference can be seen as a recurrent neural network that is parameterized by the potential functions. We follow the idea and unfold both mean field variational inference and loopy belief propagation as recurrent neural network layers that are parameterized by part scores.

The time complexity of our inference procedure is \(O(n^3)\), which is lower than the \(O(n^4)\) complexity of the exact quasi-second-order inference of Cao et al. (2017) and on par with the complexity of the approximate second-order inference of Martins and Almeida (2014).

**3.3 Learning**

Given a gold parse graph \(y^*\) of sentence \(w\), the conditional distribution over possible edge \(y_{ij}^{(edge)}\) and corresponding possible label \(y_{ij}^{(label)}\) is given by:

\[
P(y_{ij}^{(edge)}|w) = \text{softmax}(Q_{ij}^{(T)}(X_{ij}))
\]

\[
P(y_{ij}^{(label)}|w) = \text{softmax}(s_{ij}^{(label)})
\]

We define the following cross entropy losses:

\[
L_{(edge)}(\theta) = - \sum_{i,j} \log(P_{\theta}(y_{ij}^{(edge)}|w))
\]

\[
L_{(label)}(\theta) = - \sum_{i,j} \mathbb{1}(y_{ij}^{*(edge)}) \log(P_{\theta}(y_{ij}^{(label)}|w))
\]

where \(\theta\) is the parameters of our model, \(\mathbb{1}(y_{ij}^{*(edge)})\) denotes the indicator function and equals 1 when edge \((i, j)\) exists in the gold parse and 0 otherwise, and \(i, j\) ranges over all the words in the sentence. We optimize the weighted average of the two losses.

\[
\mathcal{L} = \lambda L_{(label)} + (1 - \lambda) L_{(edge)}
\]

where \(\lambda\) is a hyperparameter.

**4 Experiments**

**4.1 Hyperparameters**

We tuned the hyperparameters of our baseline model from Dozat and Manning (2018) and our second-order model on the DM development set. We followed Dozat and Manning (2018) using 100-dimensional pretrained GloVe embeddings (Pennington et al., 2014) and transformed them to be 125-dimensional. Words and lemmas appeared less than 7 times are replaced with a special unknown token. We use the same dataset split as in previous approaches (Martins and Almeida, 2014; Du et al., 2015) with 33,964 sentences in the training set, 1,692 sentences in the development set,
<table>
<thead>
<tr>
<th>Hidden Layer</th>
<th>Hidden Sizes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word/GloVe/Lemma/Char</td>
<td>100</td>
</tr>
<tr>
<td>POS</td>
<td>50</td>
</tr>
<tr>
<td>GloVe Linear</td>
<td>125</td>
</tr>
<tr>
<td>BiLSTM LSTM</td>
<td>3*600</td>
</tr>
<tr>
<td>Char LSTM</td>
<td>1*400</td>
</tr>
<tr>
<td>Unary Arc/Label</td>
<td>600</td>
</tr>
<tr>
<td>Binary Arc</td>
<td>150</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dropouts</th>
<th>Dropout Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word/GloVe/POS/Lemma</td>
<td>20%</td>
</tr>
<tr>
<td>Char LSTM (FF/recurrent)</td>
<td>33%</td>
</tr>
<tr>
<td>Char Linear</td>
<td>33%</td>
</tr>
<tr>
<td>BiLSTM (FF/recurrent)</td>
<td>45%/25%</td>
</tr>
<tr>
<td>Unary Arc/Label</td>
<td>25%/33%</td>
</tr>
<tr>
<td>Binary Arc</td>
<td>25%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Optimizer &amp; Loss</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline Interpolation (λ)</td>
<td>0.025</td>
</tr>
<tr>
<td>Second-Order Interpolation (λ)</td>
<td>0.07</td>
</tr>
<tr>
<td>Adam β₁</td>
<td>0.95</td>
</tr>
<tr>
<td>Learning rate</td>
<td>$10^{-2}$</td>
</tr>
<tr>
<td>LR decay</td>
<td>0.5</td>
</tr>
<tr>
<td>L2 regularization (MF/LBP)</td>
<td>$30^{-9}/30^{-8}$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Weight Initialization</th>
<th>Mean/Stddev</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unary weight (Eq. 1)</td>
<td>0.0%/1.0</td>
</tr>
<tr>
<td>Binary weight (Eq. 3)</td>
<td>0.0%/0.25</td>
</tr>
</tbody>
</table>

Table 1: Hyperparameter for baseline and second-order models in our experiment.

1,410 sentences in the in-domain test set and 1,849 Brown Corpus sentences in the out-of-domain test set. We additionally removed sentences longer than 60 in order to speed up training, which results in 33,916 training sentences. The final hyperparameters of our baseline and second-order model are shown in Table 1. Following Dozat and Manning (2018), we used Adam (Kingma and Ba, 2014) for optimizing our model, annealing the learning rate by 0.5 for every 10,000 steps, and switched the optimizer to AMSGrad (Reddi et al., 2018) after 5,000 steps without improvement. We trained the model for 100,000 iterations with batch sizes of 6,000 tokens and terminated training early after 10,000 iterations with no improvement on the development sets.

4.2 Main Results

We compare our model with previous state-of-the-art approaches in Table 2. Du et al. (2015) is a hybrid model. A&M is from Almeida and Martins (2015). PTS17 proposed by (Peng et al., 2017) and Basic is single task parsing while Freda3 is a multitask parser across three formalisms. WCGL18 (Wang et al., 2018) is a neural transition-based model. D&M (Dozat and Manning, 2018) is a graph-based model and "Baseline" is the first-order model from Dozat and Manning (2018) that was trained by ourselves. For our model, we used mean field variational inference and loopy belief propagation for 3 iterations.

In the basic setting, on average our model outperforms the best previous one by 1.3% on the in-domain test set and 1.3% on the out-of-domain test set. With lemma and character-based embeddings, our model leads to an average improvement of 0.3% and 0.6% over previous models. Our model also outperforms the baseline by 0.2% — —0.5% on average with different settings and test sets. Dozat and Manning (2018) found that on the PAS dataset their model cannot benefit from lemma and character-based embeddings and hence speculated that they may have approached the ceiling of the PAS F1 score. As shown in our experiments on the PAS dataset, our model cannot benefit from lemma and character-based embeddings either, but it obtains higher F1 scores, which suggests that the ceiling may not have been reached.

Note that while we do not force our parser to predict a directed acyclic graph, we found that only 0.7% of the test sentences have cycles in their parses.

4.3 Analysis

Small Training Data

To evaluate the performance of our model on smaller training data, we repeated our experiments with randomly sampled 70%, 40% and 10% of the training set. Table 3 shows the F1 scores averaged over 5 runs (each time with a new randomly sampled training subset). It can be seen that the advantage of our model over the baseline increases significantly when the training data becomes smaller. We make the following speculation to explain this observation. The BiLSTM layer in the baseline and our model is capable of capturing high-order information to some extent. However, without prior knowledge of high-order parts, it may require more training data to learn this capability than a high-order decoder. So with small training data, the baseline loses the capability of utilizing high-order information, while our model can still rely on the decoder for high-order parsing.

Performance on Different Sentence Lengths

We want to study the impact of sentence lengths on first-order parsing and our second-order parsing. We split the test sets of all the formalisms into five subsets with different sentence length ranges.
Table 2: Comparison of labeled F1 scores achieved by our model and previous state-of-the-arts. The F1 scores of Baseline and our models are averaged over 5 runs. ID denotes the in-domain (WSJ) test set and OOD denotes the out-of-domain (Brown) test set. +Char and +Lemma means augmenting the token embeddings with character-level and lemma embeddings.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>PAS</th>
<th>PSD</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ID</td>
<td>OOD</td>
<td>ID</td>
<td>OOD</td>
</tr>
<tr>
<td>Du et al. (2015)</td>
<td>89.1</td>
<td>81.8</td>
<td>91.3</td>
<td>87.2</td>
</tr>
<tr>
<td>A&amp;M, (2015)</td>
<td>88.2</td>
<td>81.8</td>
<td>90.9</td>
<td>86.9</td>
</tr>
<tr>
<td>WCGL, (2018)</td>
<td>90.3</td>
<td>84.9</td>
<td>91.7</td>
<td>87.6</td>
</tr>
<tr>
<td>PTS17: Basic</td>
<td>89.4</td>
<td>84.5</td>
<td>92.2</td>
<td>88.3</td>
</tr>
<tr>
<td>PTS17: Freda3</td>
<td>90.4</td>
<td>85.3</td>
<td>92.7</td>
<td>89.0</td>
</tr>
<tr>
<td>D&amp;M, (2018): Basic</td>
<td>91.4</td>
<td>86.9</td>
<td>93.9</td>
<td>90.8</td>
</tr>
<tr>
<td>Baseline: Basic</td>
<td>92.6</td>
<td>88.0</td>
<td>94.1</td>
<td>91.0</td>
</tr>
<tr>
<td>MF: Basic</td>
<td>93.0</td>
<td>88.4</td>
<td>94.3</td>
<td>91.5</td>
</tr>
<tr>
<td>LBP: Basic</td>
<td>92.9</td>
<td>88.4</td>
<td>94.3</td>
<td>91.5</td>
</tr>
<tr>
<td>D&amp;M, (2018): +Char +Lemma</td>
<td>93.7</td>
<td>88.9</td>
<td>93.9</td>
<td>90.6</td>
</tr>
<tr>
<td>Baseline: +Char +Lemma</td>
<td>93.7</td>
<td>89.4</td>
<td>94.1</td>
<td>90.9</td>
</tr>
<tr>
<td>MF: +Char +Lemma</td>
<td>94.0</td>
<td>89.7</td>
<td>94.1</td>
<td>91.3</td>
</tr>
<tr>
<td>LBP: +Char +Lemma</td>
<td>93.9</td>
<td>89.5</td>
<td>94.2</td>
<td>91.3</td>
</tr>
</tbody>
</table>

Table 3: Comparison of labeled F1 scores achieved by our model and our baseline on 10%, 40%, 70% of the training data. We report the average F1 score over 5 runs with different randomly sampled training data.

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>PAS</th>
<th>PSD</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ID</td>
<td>OOD</td>
<td>ID</td>
<td>OOD</td>
</tr>
<tr>
<td>Baseline: 70%</td>
<td>92.0</td>
<td>87.0</td>
<td>93.8</td>
<td>90.6</td>
</tr>
<tr>
<td>MF: 70%</td>
<td>92.4</td>
<td>87.5</td>
<td>93.9</td>
<td>90.8</td>
</tr>
<tr>
<td>LBP: 70%</td>
<td>92.3</td>
<td>87.4</td>
<td>94.0</td>
<td>90.9</td>
</tr>
<tr>
<td>Baseline: 40%</td>
<td>90.8</td>
<td>85.5</td>
<td>93.2</td>
<td>89.6</td>
</tr>
<tr>
<td>MF: 40%</td>
<td>91.2</td>
<td>86.0</td>
<td>93.4</td>
<td>90.0</td>
</tr>
<tr>
<td>LBP: 40%</td>
<td>91.2</td>
<td>86.0</td>
<td>93.5</td>
<td>90.0</td>
</tr>
<tr>
<td>Baseline: 10%</td>
<td>86.1</td>
<td>80.0</td>
<td>90.8</td>
<td>86.4</td>
</tr>
<tr>
<td>MF: 10%</td>
<td>86.9</td>
<td>81.0</td>
<td>91.3</td>
<td>87.1</td>
</tr>
<tr>
<td>LBP: 10%</td>
<td>86.8</td>
<td>80.9</td>
<td>91.3</td>
<td>87.0</td>
</tr>
</tbody>
</table>

and evaluate our model and the baseline on them. Figure 4 shows that our model has more advantage over the baseline when sentences get longer, especially when sentences are longer than 40. One possible explanation is that BiLSTM has difficulty in capturing long-range dependencies in long sentences, which leads to lower performance on the first-order baseline; but such long-range dependencies can still be captured with second-order parsing. It can also be seen that on long sentences, our model has more advantage over the baseline for the out-of-domain test set than for the in-domain test set, which suggests that our model has better generalizability especially on long sentences.

**Mean Field vs. Loopy Belief Propagation**

We compare mean field variational inference and loopy belief propagation algorithms in Table 4. We tuned the hyperparameters of our model for each algorithm and iteration number separately. We find that in general mean field variational inference has very similar performance to loopy belief propagation. In addition, with more iterations, the performance of mean field variational inference steadily increases while the at the second iteration.

**Ablation Study**

We study how different types of second-order parts defined in Section 3.1 affect the performance of our parser. We trained our model with each
Figure 4: Relative improvements over our baseline in different sentence length intervals with different training data sizes. We report the average F1 score improvements over all the formalisms with 5 runs for each.

<table>
<thead>
<tr>
<th>Iteration</th>
<th>ID OOD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>MF 92.78 88.38</td>
</tr>
<tr>
<td>2</td>
<td>MF 92.86 88.37</td>
</tr>
<tr>
<td>3</td>
<td>MF 92.98 88.44</td>
</tr>
<tr>
<td>1</td>
<td>LBP 92.88 88.29</td>
</tr>
<tr>
<td>2</td>
<td>LBP 92.84 88.17</td>
</tr>
<tr>
<td>3</td>
<td>LBP 92.88 88.36</td>
</tr>
</tbody>
</table>

Table 4: Comparison of labeled F1 scores of mean field and loopy belief propagation with different iteration numbers on the DM dataset.

<table>
<thead>
<tr>
<th></th>
<th>ID OOD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>92.60 87.98</td>
</tr>
<tr>
<td>+Siblings</td>
<td>92.85 88.31</td>
</tr>
<tr>
<td>+Co-parents</td>
<td>92.80 88.23</td>
</tr>
<tr>
<td>+Grandparents</td>
<td>92.84 88.24</td>
</tr>
</tbody>
</table>

Table 5: The performance comparison between three types of second-order parts on the DM dataset.

Table: type of second-order parts without the other two types on the DM dataset using mean field variational inference and the result is shown in Table 5. While all the three types of second-order parts can be seen to improve the parsing performance over the baseline, the sibling parts lead to the largest performance gain on both the in-domain test set and the out-of-domain test set.

4.4 Case Study

We provide a parsing example in Figure 5 to show how our second-order parser with 3 iterations of mean field variational inference works. Before the first iteration, the marginal distributions of edges $Q_{ij}$ is initialized with unary potentials and thus is exactly what a first-order parser would produce. In the subsequent iterations, the distributions are updated with binary potentials taken into account. For each version of the distributions, we can extract a parse graph by collecting edges with probabilities larger than 0.5. From Figure 5, we can see that erroneous edges are gradually fixed through iterations. Edge $(\text{were}, \text{Poles})$ sends a strong negative co-parents message to edge $(<\text{TOP}>, \text{Poles})$ in the first iteration, so the latter has a lower probability in subsequent iterations. Edge $(\text{were}, \text{Poles})$ also sends a strong positive grandparent message to edge $(<\text{TOP}>, \text{were})$ to enhance its probability, and the latter sends an increasingly positive message back to the former in subsequent iterations. In the second and third iterations, $(\text{were}, \text{Poles})$ sends positive sibling messages and $(<\text{TOP}>, \text{were})$ sends positive grandparent messages to enhance probabilities of edges $(\text{were}, \text{They})$ and $(\text{were}, \text{not})$, which finally leads to the correct parse.

4.5 Running Speed

Our model have a time complexity of $O(d_u^2 + d_b^2 + n^3)$ while the first order model of Dozat and Manning (2018) has a time complexity of $O(d_u^2 + n^2)$ in scoring and decoding (where $d_u$ and $d_b$ are the hidden sizes of the biaffine layer and trilinear layer and $n$ is the sentence length). We compare these models with respect to training speed and parsing...
They were not Poles

Figure 5: An example of message passing (left) and the corresponding graph parses (right) in our second-order parser with mean field variational inference. We regard terms in Eq. 7 as messages sent from other arcs. Blue arcs and red arcs on the left represent positive messages (which encourage the target edge to exist) and negative messages (which discourage the target edge to exist) respectively. Lightness of the arc color represents the message intensity. Blackness of each nodes represents the probability of edge existence. A Node with a double circle means the corresponding edge is predicted to exist. Messages with low intensities are omitted in the graph. Dotted arcs and red arcs on the right represent missed predictions and wrong predictions compared to the golden parse. The period in the sentence is omitted for simplicity.

<table>
<thead>
<tr>
<th>(sents/sec)</th>
<th>train</th>
<th>parse</th>
<th>long</th>
<th>short</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>730</td>
<td>904</td>
<td>334</td>
<td>1762</td>
</tr>
<tr>
<td>MF</td>
<td>472</td>
<td>722</td>
<td>240</td>
<td>1485</td>
</tr>
<tr>
<td>LBP</td>
<td>258</td>
<td>300</td>
<td>88</td>
<td>1246</td>
</tr>
</tbody>
</table>

Table 6: Training and parsing speed (sentences/second) comparison of the baseline and our model (3 iterations for our second-order parser). long means the parsing speed on sentences longer than 40 and short means the parsing speed on sentences no longer than 10.

speed on an Nvidia Tesla P40 server. The result is shown in Table 6. Mean field variational inference slows down training and parsing by 35% and 20% respectively compared with the baseline. However, loopy belief propagation slows down training and parsing by 65% and 67% respectively compared with the baseline.

4.6 Significance Test

We trained 25 basic models of our approach and the baseline with the same hyperparameters in Table 1 on each formalism. Student’s t-test shows that our second-order model outperforms our baseline model on all the formalisms with a significance level of 0.005.

5 Related Work

5.1 Semantic Dependency Parsing

Semantic dependency parsing can be classified as transition-based approaches and graph-based approaches. Wang et al. (2018) proposed a transition-based parser for semantic parsing, while Du et al. (2015) proposed a hybrid parser that benefits from both transition-based approaches and graph-based approaches. Peng et al. (2017) proposed a graph-based approach that trains on all the three formalisms simultaneously and Peng
et al. (2018) further proposed to learn from different corpora. Dozat and Manning (2018) proposed a graph-based simple but powerful neural network for semantic dependency parsing using a bilinear or biaffine (Dozat and Manning, 2016) layer to encode the interaction between words. Most of these approaches proposed first-order dependency parser while Martins and Almeida (2014) proposed a way to encode higher-order parts with hand-crafted features and introduced a novel co-parent part for semantic dependency parsing. They used discrete optimizing algorithm alternating directions dual decomposition (AD3) as their decoder. Cao et al. (2017) also proposed a quasi-second-order semantic dependency parser with dynamic programming. Our model contains second-order information comparing with the first-order approaches and benefits from end-to-end training comparing with other second-order approaches.

5.2 Higher-Order Dependency Parsing

Higher-order parsing has been extensively studied in the literature of syntactic dependency parsing. Much of these work is based on the first-order maximum spanning tree (MST) parser of McDonald et al. (2005) which factorizes a dependency tree into individual edges and maximizes the summation of the scores of all the edges in a tree. McDonald and Pereira (2006) introduced a second-order MST that factorizes a dependency tree into not only edges but also second-order sibling parts, which allows interactions between adjacent sibling words. Carreras (2007) defined second-order grandparent parts representing grandparental relations. Koo and Collins (2010) introduced third-order grand-sibling and tri-sibling parts. A grand-sibling part represents a grandparent with two grandchildren and a tri-sibling part represents a parent with three children. Ma and Zhao (2012) defined grand-tri-sibling parts for fourth-order dependency parsing.

Many previous approaches to higher-order dependency parsing perform exact decoding based on dynamic programming, but there is also research in approximate higher-order parsing. Martins et al. (2011) proposed an alternating directions dual decomposition (AD3) algorithm which splits the original problem into several local subproblems and solves them iteratively. They employed AD3 for second-order dependency parsing to speed up decoding. Smith and Eisner (2008) and Gormley et al. (2015) proposed to use belief propagation for approximate higher-order parsing, which is closely related to our work. While higher-order parsing has been shown to improve syntactic dependency parsing accuracy, it receives less attention in semantic dependency parsing. Martins and Almeida (2014) proposed second-order semantic dependency parsing and employed AD3 for approximate decoding. Cao et al. (2017) proposed a quasi-second-order parser and used dynamic programming for decoding with time complexity of $O(n^4)$.

5.3 CRF as Recurrent Neural Networks

Zheng et al. (2015) are probably the first to propose the idea of unfolding iterative inference algorithms on CRFs as a stack of recurrent neural network layers. They unfolded mean field variational inference in a neural network designed for semantic segmentation. There is a lot of subsequent work that employs this technique, especially in the computer vision area. For example, Zhu et al. (2017) proposed a structured attention neural model for Visual Question Answering with a CRF over image regions and unfolded both mean field variational inference and loopy belief propagation algorithms as recurrent layers.

6 Conclusion

We proposed a novel graph-based second-order parser for semantic dependency parsing. We constructed an end-to-end neural network that uses a trilinear function to score second-order parts and finds the highest-scoring parse graph by either mean field variational inference or loopy belief propagation algorithms unfolded as recurrent neural network layers. Our experimental results show that our model outperforms previous state-of-the-art model and has higher accuracies especially on out-of-domain data and long sentences. Our code is publicly available at https://github.com/wangxinyu0922/Second_Order_SDP

Acknowledgments

This work was supported by the Major Program of Science and Technology Commission Shanghai Municipal (17JC1404102).
References


