Transition-Based Parsing for Deep Dependency Structures

Xun Zhang *
Peking University

Yantao Du *
Peking University

Weiwei Sun *
Peking University

Xiaojun Wan *
Peking University

Derivations under different grammar formalisms allow extraction of various dependency structures. Particularly, bilexical deep dependency structures beyond surface tree representation can be derived from linguistic analysis grounded by CCG, LFG, and HPSG. Traditionally, these dependency structures are obtained as a by-product of grammar-guided parsers. In this article, we study the alternative data-driven, transition-based approach, which has achieved great success for tree parsing, to build general dependency graphs. We integrate existing tree parsing techniques and present two new transition systems that can generate arbitrary directed graphs in an incremental manner. Statistical parsers that are competitive in both accuracy and efficiency can be built upon these transition systems. Furthermore, the heterogeneous design of transition systems yields diversity of the corresponding parsing models and thus greatly benefits parser ensemble. Concerning the disambiguation problem, we introduce two new techniques, namely, transition combination and tree approximation, to improve parsing quality. Transition combination makes every action performed by a parser significantly change configurations. Therefore, more distinct features can be extracted for statistical disambiguation. With the same goal of extracting informative features, tree approximation induces tree backbones from dependency graphs and re-uses tree parsing techniques to produce tree-related features. We conduct experiments on CCG-grounded functor–argument analysis, LFG-grounded grammatical relation analysis, and HPSG-grounded semantic dependency analysis for English and Chinese. Experiments demonstrate that data-driven models with appropriate transition systems can produce high-quality deep dependency analysis, comparable to more complex grammar-driven
models. Experiments also indicate the effectiveness of the heterogeneous design of transition systems for parser ensemble, transition combination, as well as tree approximation for statistical disambiguation.

1. Introduction

The derivations licensed by a grammar under deep grammar formalisms, for example, combinatory categorial grammar (CCG; Steedman 2000), lexical-functional grammar (LFG; Bresnan and Kaplan 1982) and head-driven phrase structure grammar (HPSG; Pollard and Sag 1994), are able to produce rich linguistic information encoded as bilexical dependencies. Under CCG, this is done by relating the lexical heads of functor categories and their arguments (Clark, Hockenmaier, and Steedman 2002). Under LFG, bilexical grammatical relations can be easily derived as the backbone of F-structures (Sun et al. 2014). Under HPSG, predicate–argument structures (Miyao, Ninomiya, and ichi Tsujii 2004) or reduction of minimal recursion semantics (Ivanova et al. 2012) can be extracted from typed feature structures corresponding to whole sentences. Dependency analysis grounded in deep grammar formalisms is usually beyond tree representations and well-suited for producing meaning representations. Figure 1 is an example from CCGBank. The deep dependency graph conveniently represents more semantically motivated information than the surface tree. For instance, it directly captures the Agent–Predicate relations between word “people” and conjuncts “fight,” “eat,” as well as “drink.”

Automatically building deep dependency structures is desirable for many practical NLP applications, for example, information extraction (Miyao et al. 2008) and question answering (Reddy, Lapata, and Steedman 2014). Traditionally, deep dependency graphs are generated as a by-product of grammar-guided parsers. The challenge is that a deep-grammar-guided parsing model usually cannot produce full coverage and the time complexity of the corresponding parsing algorithms is very high. Previous work on data-driven dependency parsing mainly focused on tree-shaped representations. Nevertheless, recent work has shown that a data-driven approach is also applicable to generate more general linguistic graphs. Sagae and Tsujii (2008) present an initial study on applying transition-based methods to generate HPSG-style predicate–argument structures, and have obtained competitive results. Furthermore, Titov et al. (2009) and Henderson et al. (2013) have shown that more general graphs rather than planars can be produced by augmenting existing transition systems.

This work follows early encouraging research and studies transition-based approaches to construct deep dependency graphs. The computational challenge to incremental graph spanning is the existence of a large number of crossing arcs in deep

Figure 1
An example from CCGBank. The upper curves represent a deep dependency graph and the bottom curves represent a traditional dependency tree.
dependency analysis. To tackle this problem, we integrate insightful ideas, especially the ones illustrated in Nivre (2009) and Gómez-Rodríguez and Nivre (2010), developed in the tree spanning scenario, and design two new transition systems, both of which are able to produce arbitrary directed graphs. In particular, we explore two techniques to localize transition actions to maximize the effect of a greedy search procedure. In this way, the corresponding parsers for generating linguistically motivated bilexical graphs can process sentences in close to linear time with respect to the number of input words. This efficiency advantage allows deep linguistic processing for very-large-scale text data.

For syntactic parsing, ensembled methods have been shown to be very helpful in boosting accuracy (Sagae and Lavie 2006; Zhang et al. 2009; McDonald and Nivre 2011). In particular, Surdeanu and Manning (2010) presented a nice comparative study on various ensemble models for dependency tree parsing. They found that the diversity of base parsers is more important than complex ensemble models for learning. Motivated by this observation, the authors proposed a hybrid transition-based parser that achieved state-of-the-art performance by combining complementary prediction powers of different transition systems. One advantage of their architecture is the linear-time decoding complexity, given that all base models run in linear-time. Another concern of our work is about the model diversity obtained by the heterogeneous design of transition systems for general graph spanning. Empirical evaluation indicates that statistical parsers built upon our new transition systems as well as the existing best transition system—namely, Titov et al. (2009)’s system (THMM, hereafter)—exhibit complementary parsing strengths, which benefit system combination. In order to take advantage of this model diversity, we propose a simple yet effective ensemble model to build a better hybrid system.

We implement statistical parsers using the structured perceptron algorithm (Collins 2002) for transition classification and use a beam decoder for global inference. Concerning the disambiguation problem, we introduce two new techniques, namely, transition combination and tree approximation, to improve parsing quality. To increase system coverage, the ARC transitions designed by the THMM as well as our systems do not change the nodes in the stack nor buffer in a configuration: Only the nodes linked to the top of the stack or buffer are modified. Therefore, features derived from the configurations before and after an ARC transition are not distinct enough to train a good classifier. To deal with this problem, we propose the transition combination technique and three algorithms to derive oracles for modified transition systems. When we apply our models to semantics-oriented deep dependency structures, for example, CCG-grounded functor–argument analysis and HPSG-grounded reduced minimal recursion semantics (MRS; Copestake et al. 2005) analysis, we find that syntactic trees can provide very helpful features. In case the syntactic information is not available, we introduce a tree approximation technique to induce tree backbones from deep dependency graphs. Such tree backbones can be utilized to train a tree parser which provides pseudo tree features.

To evaluate transition-based models for deep dependency parsing, we conduct experiments on CCG-grounded functor–argument analysis (Hockenmaier and Steedman 2007; Tse and Curran 2010), LFG-grounded grammatical relation analysis (Sun et al. 2014), and HPSG-grounded semantic dependency analysis (Miyao, Ninomiya, and ichi Tsujii 2004; Ivanova et al. 2012) for English and Chinese. Empirical evaluation indicates some non-obvious facts:

1. Data-driven models with appropriate transition systems and disambiguation techniques can produce high-quality deep dependency analysis, comparable to more complex grammar-driven models.
2. Parsers built upon heterogeneous transition systems and decoding orders have complementary prediction strengths, and the parsing quality can be significantly improved by system combination; compared to the best individual system, system combination gets an absolute labeled F-score improvement of 1.21 on average.

3. Transition combination significantly improves parsing accuracy on a wide range of conditions, resulting in an absolute labeled F-score improvement of 0.74 on average.

4. Pseudo trees contribute to semantic dependency parsing (SDP) equally well to syntactic trees, and result in an absolute labeled F-score improvement of 1.27 on average.

We compare our parser with representative state-of-the-art parsers (Miyao and Tsujii 2008; Auli and Lopez 2011b; Martins and Almeida 2014; Xu, Clark, and Zhang 2014; Du, Sun, and Wan 2015) with respect to different architectures. To evaluate the impact of grammatical knowledge, we compare our parser with parsers guided by treetbank-induced HPSG and CCG grammars. Both of our individual and ensembled parsers achieve equivalent accuracy to HPSG and CCG chart parsers (Miyao and Tsujii 2008; Auli and Lopez 2011b), and outperform a shift-reduce CCG parser (Xu, Clark, and Zhang 2014). It is worth noting that our parsers exclude all syntactic and grammatical information. In other words, strictly less information is used. This result demonstrates the effectiveness of data-driven approaches to the deep linguistic processing problem. Compared to other types of data-driven parsers, our individual parser achieves equivalent performance to and our hybrid parser obtains slightly better results than factorization parsers based on dual decomposition (Martins and Almeida 2014; Du, Sun, and Wan 2015). This result highlights the effectiveness of the lightweight, transition-based approach.

Parsers based on the two new transition systems have been utilized as base components for parser ensemble (Du et al. 2014) for SemEval 2014 Task 8 (Oepen et al. 2014). Our hybrid system obtained the best overall performance of the closed track of this shared task. In this article, we re-implement all models, calibrate features more carefully, and thus obtain improved accuracy. The idea to extract tree-shaped backbone from a deep dependency graph has also been used to design other types of parsing models in our early work (Du et al. 2014, 2015; Du, Sun, and Wan 2015). Nevertheless, the idea to train a pseudo tree parser to serve a transition-based graph parser is new.

The implementation of our parser is available at http://www.icst.pku.edu.cn/lcwm/grass.

2. Transition Systems for Graph Spanning

2.1 Background Notations

A dependency graph \( G = (V, A) \) is a labeled directed graph, such that for sentence \( x = w_1, \ldots, w_n \) the following holds:

1. \( V = \{0, 1, 2, \ldots, n\} \),
2. \( A \subseteq V \times \mathbb{R} \times V \).
The vertex set $V$ consists of $n + 1$ nodes, each of which is represented by a single integer. In particular, 0 represents a virtual root node $w_0$, and all others correspond to words in $x$. The arc set $A$ represents the labeled dependency relations of the particular analysis $G$. Specifically, an arc $(i, r, j) \in A$ represents a dependency relation $r$ from head $w_i$ to dependent $w_j$. A dependency graph $G$ is thus a set of labeled dependency relations between the root and the words of $x$. To simplify the description in this section, we mainly consider unlabeled parsing and assume the relation set $R$ is a singleton. Or, taking it another way, we assume $A \subseteq V \times V$. It is straightforward to adapt the discussions in this article for labeled parsing. To do so, we can parameterize transitions with possible dependency relations. For empirical evaluation as discussed in Section 5, we will test both labeled and unlabeled parsing models.

Following Nivre (2008), we define a transition system for dependency parsing as a quadruple $S = (C, T, cs, C_t)$, where:

1. $C$ is a set of configurations, each of which contains a buffer $\beta$ of (remaining) words and a set $A$ of dependency arcs,
2. $T$ is a set of transitions, each of which is a (partial) function $t : C \mapsto C$,
3. $cs$ is an initialization function, mapping a sentence $x$ to a configuration with $\beta = [1, \ldots, n]$,
4. $C_t \subseteq C$ is a set of terminal configurations.

Given a sentence $x = w_1, \ldots, w_n$ and a graph $G = (V, A)$ on it, if there is a sequence of transitions $t_1, \ldots, t_m$ and a sequence of configurations $c_0, \ldots, c_m$ such that $c_0 = cs(x)$, $t_i(c_{i-1}) = c_i$ ($i = 1, \ldots, m$), $c_m \in C_t$, and $A_{c_m} = A$, we say the sequence of transitions is an oracle sequence. And we define $\bar{A}_{c_i} = A - A_{c_i}$ for the arcs to be built of $c_i$. We could denote a transition sequence as either $t_1, \ldots, t_m$ or $c_0, \ldots, c_m$.

In a typical transition-based parsing process, the input words are put into a queue and partially built structures are organized by a stack. A set of SHIFT/REDUCE actions are performed sequentially to consume words from the queue and update the partial parsing results organized by the stack. Our new systems designed for deep parsing differ with respect to their information structures to define a configuration and the behaviors of transition actions.

### 2.2 Naive Spanning and Locality

For every two nodes, a simple graph-spanning strategy is to check if they can be directly connected by an arc. Accordingly, a “naive” spanning algorithm can be implemented by exploring a left-to-right checking order, as introduced by Covington (2001) and modified by Nivre (2008).

\[
\text{PARSE}(x = (w_1, \ldots, w_n))
\]

1. for $j = 1..n$
2. for $k = j - 1..1$
3. Link($j, k$

The operation $\text{Link}$ chooses between 1) adding the arc $(i, j)$ or $(j, i)$ and 2) adding no arc at all. In this way, the algorithm builds a graph by incrementally trying to link every pair of words.
The complexity of naive spanning is $\theta(n^2)$, because it does nothing to explore the topological properties of a linguistic structure. In other words, the naive graph-spanning idea does not fully take advantages of the greedy search of the transition-based parsing architecture. On the contrary, a well-designed transition system for (projective) tree parsing can decode in linear time by exploiting locality among subtrees. Take the arc-eager system presented in Nivre (2008), for example: Only the nodes at the top of the stack and the buffer are allowed to be linked. Such limitation is the key to implement a linear time decoder. In the following, we introduce two ideas to localize a transition action, that is, to allow a transition to manipulate only the frontier items in the data structures of a configuration. By this means, we can decrease the number of possible transitions for each configuration and thus minimize the total decoding time.

### 2.3 System 1: Online Re-ordering

The online re-ordering approach that we explore is to provide the system with ability to re-order the nodes during parsing in an online fashion. The key idea, as introduced in Titov et al. (2009) and Nivre (2009), is to allow a SWAP transition that switches the position of the two topmost nodes on the stack. By changing the linear order of words, the system is able to build crossing arcs for graph spanning. We refer to this approach as online re-ordering. We introduce a stack-based transition system with online re-ordering for deep dependency parsing. The obtained oracle parser is complete with respect to the class of all directed graphs without self-loop.

#### 2.3.1 The System

We define a transition system $S_S = (C, T, c_0, C_t)$, where a configuration $c = (\sigma, \beta, A) \in C$ contains a stack $\sigma$ of nodes, besides $\beta$ and $A$. We set the initial configuration for a sentence $x = w_1, \ldots, w_n$ to be $c_s(x) = ([1], [1, \ldots, n], \{\})$, and take $C_t$ to be the set of all configurations of the form $c_t = (\sigma, [], A)$ (for any $\sigma$ any $A$). These transitions are shown in Figure 2 and explained as follows.

- **SHIFT** (sh) removes the front from the buffer and pushes it onto the stack.
- **LEFT / RIGHT-ARC** (la/ra) updates a configuration by adding $(j, i)/(i, j)$ to $A$ where $i$ is the top of the stack, and $j$ is the front of the buffer.
- **POP** (pop) updates a configuration by popping the top of the stack.
- **SWAP** (sw) updates a configuration with stack $\sigma|ij$ by moving $i$ back to the buffer.

---

1 We assume that at most one edge exists between two words. This is a reasonable assumption for a linguistic representation.
A variation of transition \textsc{swap} is \textsc{swap}_T, which updates the configuration by swapping \(i\) and \(j\). However, the system of this variation is not complete with respect to directed graphs because the power of transition \textsc{swap}_T is limited, and counterexamples of completeness can be found. For more theoretical discussion about this system (i.e., THMM), see Titov et al. (2009). We also denote Titov et al. (2009)'s system as \(S_T\).

2.3.2 Theoretical Analysis. The soundness of \(S_S\) is trivial. To demonstrate the completeness of the system, we give a constructive proof that can derive oracle transitions for any arbitrary graph. To simplify the description, the label attached to transitions are not considered. The idea is inspired by Titov et al. (2009). Given a sentence \(x = w_1, \ldots, w_n\), and a graph \(G = (V,A)\) on it, we start with the initial configuration \(c_0 = c_s(x)\) and compute the oracle transitions step by step. On the \(i\)-th step, let \(p\) be the top of \(\sigma_{c_{i-1}}\), \(b\) be the front of \(\beta_{c_{i-1}}\); let \(L(j)\) be the ordered list of nodes connected to \(j\) in \(\bar{A}_{c_{i-1}}\) for any node \(j \in \sigma_{c_{i-1}}\); let \(L(\sigma_{c_{i-1}}) = [L(j_0), \ldots, L(j_l)]\) if \(\sigma_{c_{i-1}} = [j_l, \ldots, j_0]\).

The oracle transition for each configuration is derived as follows. If there is no arc linked to \(p\) in \(\bar{A}_{c_{i-1}}\), then we set \(t_i\) to \text{pop}; if there exists \(a \in \bar{A}_{c_{i-1}}\) linking \(p\) and \(b\), then we set \(t_i\) to \text{ia} or \text{ra} correspondingly. When there are only \text{sh} and \text{sw} left, we see if there is any node \(q\) under the top of \(\sigma_{c_{i-1}}\) such that \(L(q)\) precedes \(L(p)\) by the lexicographical order. If so, we set \(t_i\) to \text{sw}; else we set \(t_i\) to \text{sh}. An example for when to do \text{sw} is shown in Figure 3. Let \(c_1 = t_i(c_{i-1})\); we continue to compute \(t_{i+1}\), until \(\beta_{c_i}\) is empty.

**Lemma 1**
If \(t_i\) is \text{sh}, \(L(\sigma_{c_{i-1}}) = [L(j_0), \ldots, L(j_l)]\) is complete ordered by lexicographical order.

**Proof**
It cannot be the case that for some \(u > 0\), \(L(j_u)\) strictly precedes \(L(j_0)\), otherwise \(t_i\) should be \text{sw}. It also cannot be the case that for some \(u > v > 0\), \(L(j_u)\) strictly precedes \(L(j_v)\), because when \(j_{v-1}\) is shifted onto the stack, \(L(j_v)\) precedes \(L(j_u)\) and all the transitions do not change \(L(j_v)\) and \(L(j_u)\) afterwards. ■

**Lemma 2**
For \(i = 0, \ldots, m\), there is no arc \((j,k) \in \bar{A}_{c_i}\) such that \(j,k \in \sigma_i\).

**Proof**
When \(j \in \sigma_{c_i}\) is shifted onto the stack by the \(w\)-th transition \(t_{iw}\), there must be no arc \((j,k)\) or \((k,j)\) in \(\bar{A}_{c_w}\) such that \(k \in \sigma_{c_w}\). Otherwise, by induction every node in \(\sigma_{c_{w-1}}\) can only link to nodes in \(\beta_{c_{w-1}}\), which implies that \(L(k)\) has one of the smallest lexicographical orders, and from Lemma 1 the top of \(\sigma_{c_{w-1}}\) must be linked to \(j\). And not \text{sh}, but \text{la} or \text{ra} should be applied. ■

![Figure 3](https://example.com/fig3.png)

**Figure 3**
\(\sigma, \beta, \text{and } \bar{A}\) of two configurations \(c_1\) and \(c_2\). In the left graphic, \(L(\sigma_{c_1}) = [[6],[5],[5,6],[7]]\).

Because \([5,6]\) and \([5]\) precedes \([6]\), we apply two \textsc{swaps} and then two \textsc{shifts}, obtaining the right graphic.
Theorem 1

$t_1, \ldots, t_m$ is an oracle sequence of transitions for $G$.

Proof

From Lemma 2, we can infer that $\bar{A}_{cm} = \emptyset$, so it suffices to show the sequence of transitions is always finite. We define a swap sequence to be a subsequence $t_i, \ldots, t_j$ such that $t_i$ and $t_j$ are sw, $t_{i-1}$ and $t_{j+1}$ are not sw, and a shift sequence similarly. It can be seen that a swap sequence is always followed by a shift sequence, the length of which is no less than the swap sequence, and if the two sequences are of the same length, the next transition cannot be sw. Let $\#(t)$ to be the number of transition types $t$ in the sequence, then $\#(la), \#(ra), \#(pop)$, and $\#(sh) - \#(sw)$ are all finite. Therefore the number of swap sequence is finite, indicating that the transition sequence is finite. ■

2.4 System 2: Two-Stack–Based System

A majority of transition systems organize partial parsing results with a stack. Classical parsers, including arc-standard and arc-eager ones, add dependency arcs only between nodes that are adjacent on the stack or the buffer. A natural idea to produce crossing arcs is to temporarily move nodes that block non-adjacent nodes to an extra memory module, like the two-stack–based system for two-planar graphs (Gómez-Rodríguez and Nivre 2010) and the list-based system (Nivre 2008). In this article, we design a new transition system to handle crossing arcs by using two stacks. This system is also complete with respect to the class of directed graphs without self-loop.

2.4.1 The System. We define the two-stack–based transition system $S_{2S} = (C, T, c_s, C_t)$, where a configuration $c = (\sigma, \sigma', \beta, A) \in C$ contains a primary stack $\sigma$ and a secondary stack $\sigma'$. We set $c_s(x) = ([], [1, \ldots, n], \{\})$ for the sentence $x = w_1, \ldots, w_n$, and we take the set $C_t$ to be the set of all configurations with empty buffers. The transition set $T$ contains six types of transitions, as shown in Figure 4. We only explain MEM and RECALL:

- MEM (mem) pops the top element from the primary stack and pushes it onto the secondary stack.
- RECALL (rc) moves the top element of the secondary stack back to the primary stack.

<table>
<thead>
<tr>
<th>Transition</th>
<th>Configuration</th>
</tr>
</thead>
<tbody>
<tr>
<td>LEFT-ARC</td>
<td>$(\sigma</td>
</tr>
<tr>
<td>RIGHT-ARC</td>
<td>$(\sigma</td>
</tr>
<tr>
<td>SHIFT</td>
<td>$(\sigma, \sigma', j</td>
</tr>
<tr>
<td>POP</td>
<td>$(\sigma</td>
</tr>
<tr>
<td>MEM</td>
<td>$(\sigma</td>
</tr>
<tr>
<td>RECALL</td>
<td>$(\sigma, \sigma'</td>
</tr>
</tbody>
</table>

Figure 4

Transitions of the two-stack–based system.
2.4.2 Theoretical Analysis. The soundness of this system is trivial, and the completeness is also straightforward after we give the construction of an oracle transition sequence for an arbitrary graph. The oracle is computed as follows on the i-th step: We do la, ra, and pop transitions just like in Section 2.3.2. After that, let b be the front of $\beta_{ci-1}$, we see if there is $j \in \sigma_{ci-1}$ or $j \in \sigma'_{ci-1}$ linked to b by an arc in $\bar{A}_{ci-1}$. If $j \in \sigma_{ci-1}$, then we do a sequence of mem to make j the top of $\sigma_{ci-1}$; if $j \in \sigma'_{ci-1}$, then we do a sequence of rc to make j the top of $\sigma'_{ci-1}$. When no node in $\sigma_{ci-1}$ or $\sigma'_{ci-1}$ is linked to b, we do sh.

Theorem 2
$S_{2S}$ is complete with respect to directed graphs without self-loop.

Proof
The completeness immediately follows the fact that the computed oracle sequence is finite, and every time a node is shifted onto $\sigma_{ci}$, no arc in $\bar{A}_{ci}$ links nodes in $\sigma_{ci}$ to the shifted node. ■

2.4.3 Related Systems. Gómez-Rodríguez and Nivre (2010, 2013) introduced a two-stack–based transition system for tree parsing. Their study is motivated by the observation that the majority of dependency trees in various treebanks are actually planar or two-planar graphs. Accordingly, their algorithm is specially designed to handle projective trees and two-planar trees, but not all graphs. Because many more crossing arcs exist in deep dependency structures and more sentences are assigned with neither planar nor two-planar graphs, their strategy of utilizing two stacks is not suitable for the deep dependency parsing problem. Different from their system, our new system maximizes the utility of two memory modules and is able to handle any directed graphs.

The list-based systems, such as the basic one introduced by Nivre (2008) and the extended one introduced by Choi and Palmer (2011), also use two memory modules. The function of the secondary memory module of their systems and ours is very different. In our design, only nodes involved in a subgraph that contains crossing arcs may be put into the second stack. In the existing list-based systems, both lists are heavily used, and nodes may be transferred between them many times. The function of the two lists is to simulate one memory module that allows accessing any unit in it.

2.5 Extension

2.5.1 Graphs with Loops. It is easy to extend our system to generate arbitrary directed graphs by adding a new transition:

- SELF-ARC adds an arc from the top element of the primary memory module ($\sigma$) to itself, but does not update any stack nor buffer.

Theorem 3
$S_S$ and $S_{2S}$ augmented with SELF-ARC are complete with respect to directed graphs.

2.5.2 Labeled Parsing and Supertagging. It is also straightforward to adapt the two-transition systems for labeled dependency graph generation. To do so, we can parameterize LEFT-ARC and RIGHT-ARC transitions with dependency relations. For example, a parameterized transition LEFT-ARC, tells the system not only that there is an arc between the frontier node of the stack and the frontier node of the buffer but also that this arc holds a relation $r$. Some linguistic representations assign labels to nodes as well. When a
deep grammar is considered to license to representation, node labels are usually called “supertags.” To assign supertags to words, namely, nodes in a dependency graph, we can parameterize the SHIFT transition with tag labels.

3. Statistical Disambiguation

3.1 Transition Classification

A transition-based parser must decide which transition is appropriate given its parsing environment (i.e., configuration). As with many other data-driven dependency parsers, we use a global linear model for disambiguation. In other words, a discriminative classifier is utilized to approximate the oracle function for a transition system $S$ that maps a configuration $c$ to a transition $t$ that is defined on $c$. More formally, a transition-based statistical parser tries to find the transition sequence $c_{0,m}$ that maximizes the following score

$$\text{SCORE}(c_{0,m}) = \sum_{i=0}^{m-1} \text{SCORE}(c_i, t_{i+1})$$

Following the state-of-the-art discriminative disambiguation technique for data-driven parsing, we define the score function as a linear combination of features defined over a configuration and a transition, as follows:

$$\text{SCORE}(c_i, t_{i+1}) = \theta^\top \phi(c_i, t_{i+1})$$

where $\phi$ defines a vector for each configuration–transition pair and $\theta$ is the weight vector for linear combination.

Exact calculation of the maximization is extremely hard without any assumption of $\phi$. Even with a proper $\phi$ for real-word parsing, exact decoding is still impractical for most practical feature designs. In this article, we follow the recent success of using beam search for approximate decoding. During parsing, the parser keeps track of multiple yet a fixed number of partial outputs to avoid making decisions too early. Training a parser in the discriminative setting corresponds to estimating $\theta$ associated with rich features. Previous research on dependency parsing shows that structured perceptron (Collins 2002; Collins and Roark 2004) is one of the strongest learning algorithms. In all experiments, we use the averaged perceptron algorithm with early update to estimate parameters. The whole parser is very similar to the transition-based system introduced in Zhang and Clark (2008, 2011b).

3.2 Transition Combination

In either THMM, $S_S$, or $S_{2S}$, the LEFT/RIGHT-ARC transition does not modify either the stack or the buffer. Only new edges are added to the target graph. When automatic classifiers are utilized to approximate an oracle, a majority of features for predicting an ARC transition will be overlapped with the features for the successive transition. Empirically, this property significantly decreases the parsing accuracy. A key observation of a linguistically motivated bilexical graph is that there is usually at most one edge between any two words, therefore an ARC transition is not followed by another ARC. As a result, any ARC with its successive transition modifies a configuration much. To practically
improve the performance of a statistical parser, we combine every pair of two successive transitions starting with ARC and transform the proposed two transition systems into two modified ones. For example, in our two-stack–based system, after combining, we obtain the transitions presented in Figure 5.

The number of edges between any two words could be at most two in real data. If there are two edges between two words \( w_a \) and \( w_b \), it must be \( w_a \rightarrow w_b \) and \( w_b \rightarrow w_a \). We call these two edges a two-cycle, and call this problem the two-cycle problem. In our combined transitions, a LEFT/RIGHT-ARC transition should appear before a non-ARC transition. In order to generate two edges between two words, we have two strategies:

A) Add a new type of transitions to each system, which consist of a LEFT-ARC transition, a RIGHT-ARC transition, and any other non-ARC transition (e.g., LEFT-ARC-RIGHT-ARC-RECALL for \( S_{25} \)).

B) Use a non-directional ARC transition instead of LEFT/RIGHT-ARC. Here, an ARC transition may add one or two edges depends on its label. In detail, we propose two algorithms, namely, ENCODELABEL and DECODELABEL (see Algorithms 1 and 2), to deal with labels for ARC transition.

### Algorithm 1 encode label

```plaintext
1: procedure ENCODELABEL(type, lLabel, rLabel)
2:     if type == LEFT then
3:         return "left" + lLabel
4:     else if type == RIGHT then
5:         return "right" + rLabel
6:     else
7:         return "both" + lLabel + "+" + rLabel
8: end if
9: end procedure
```

![Figure 5](image-url)

Original and combined transitions for the two-stack combined system. Two-cycle is not considered here.
### Algorithm 2 decode combined label, return a pair of left label and right label

1: **procedure** `DECODELABEL(label)`

2:    **if** `label.startswith("left")` **then**
3:        **return** `{label[4 :], nil}`
4:    **else if** `label.startswith("right")` **then**
5:        **return** `{nil, label[5 :]}`
6:    **else**
7:        **return** `{label[4 : label.index('|')], label[(label.index('|') + 1) :]}`
8: **end if**
9: **end procedure**

To our best efforts, the strategy B performs better.

First, let us consider accuracy. Generally speaking, it is harder for transition classification if more target transitions are defined. Using strategy A, we should add additional transitions to handle the two-cycle condition. Based on our experiments, the performance decreases when using more transitions.

Considering efficiency, we can save time by only using labels that appear in training data in strategy B. If we have a total of $K$ possible labels in training data, they will generate $K^2$ two-cycle types, but only $k$ possible combinations of two-cycle appear in training data ($k \ll K^2$). In strategy A, we must add $K^2$ transitions to deal with all possible two-cycle types, but most of them do not make sense. Using fewer two-cycle types helps us eliminate the invalid calculation and save time effectively.

Using strategy B, we change the original edges’ labels and use the `ARC(label)`–non-ARC transition instead of LEFT/RIGHT-ARC(label)–non-ARC. An `ARC(label)`–non-ARC transition should execute the `ARC(label)` transition first, then execute the non-ARC transition. `ARC(label)` generates one or two edges depends on its label. Not only do we encode two-cycle labels, but also LEFT/RIGHT-ARC labels. In practice, we only use those labels that appear in training data. Because labels that do not appear only contribute non-negative weights while training, we can eliminate them without any performance loss.

For each transition system and each dependency graph, we generate an oracle transition, and train our model according to this oracle. The constructive proofs presented in Section 2.3 and 2.4 define two kinds of oracles. However, they are not directly applicable when the transition combination strategy is utilized. The main challenge is the existence of cycles. In this article, we propose three algorithms to derive oracles for THMM, $S_S$, and $S_{2S}$, respectively. Algorithms 3 to 5 illustrate the key steps of the procedure of our system, which find the next transition $t$ given a configuration $c$ and gold graph $G_{gold} = (V_x, A_{gold})$ for the three systems. When this key procedure, namely, the `EXTRACTONEORACLE` method, is well defined, the entire transition system can be derived as follows:

```
EXTRACTORACLE(c₀, A_{gold})
1   oracle = ∅
2   while $t \leftarrow$ EXTRACTONEORACLE($c₀, A_{gold}, nil$) do
3       oracle.push_back = t
4       $c₀ \leftarrow t(c₀)$
5   end while
```
Algorithm 3 Oracle generation for the THMM system

1: procedure EXTRACT ONE ORACLE (c, A\text{gold}, label)
2: if c = (σ|i, j|β, A) \land \neg\exists k [k \geq j \land \exists ![i, l, k] \in A_\text{gold}] then
3: if label = nil then
4: return REDUCE
5: else
6: return ARC(label) \circ REDUCE
7: end if
8: else if c = (σ|i, j|β, A) \land \exists ![i, l, j] \in A_\text{gold} then
9: A_\text{gold} \leftarrow A_\text{gold} - ![i, l, j]
10: return EXTRACT ONE ORACLE (c, A_\text{gold}, label)
11: else if c = (σ|i_1\ldots i_0, j|β, A) \land 0 \leq k_0 \leq j \land \exists ![i_0, l_0, k_0] \in A_\text{gold}] \land \exists ![i_1, l_1, k_1] \in A_\text{gold}] \land \neg\exists k_0' [k_0' < k_0 \land \exists ![i_0, l_0', k_0'] \in A_\text{gold}] \land \exists ![i_1, l_1', k_1] \in A_\text{gold}] \land k_0 < k_1 \lor \neg\exists k_1 [k_1 \geq j \land \exists ![i_1, l_1, k_1] \in A_\text{gold}] then
12: if label = nil then
13: return SWAP
14: else
15: return ARC(label) \circ SWAP
16: end if
17: end if
18: if c = (σ, j|β, A) then
19: if label = nil then
20: return SHIFT
21: else
22: return ARC(label) \circ SHIFT
23: end if
24: end if
25: return nil
26: end procedure

We want to emphasize that, although the EXTRACT ORACLE methods initialize the parameter LABEL in EXTRACT ONE ORACLE as nil, if an arc transition is predicted in the EXTRACT ONE ORACLE method, it will call EXTRACT ONE ORACLE recursively to return an ARC(label)–non-ARC transition and assign a value for that LABEL.

3.3 Feature Design

Developing features has been shown to be crucial to advancing the state-of-the-art in dependency parsing (Koo and Collins 2010; Zhang and Nivre 2011). To build accurate deep dependency parsers, we utilize a large set of features for transition classification.

To conveniently define all features, we use the following notation. In a configuration with stack σ and buffer β, we denote the top two nodes in σ by σ_0 and σ_1, and the front of β by β_0. In a configuration of the two-stack–based system with the second stack σ', the top element of σ' is denoted by σ'_0 and the front of β by β_0. The left-most dependent of node n is denoted by n.lc, the right-most one by n.rc. The left-most parent of node n is denoted by n.lp, the right-most one by n,rp. Then we denote the word and POS-tag
of node $n$ by $w_n, p_n$, respectively. Our parser derives the so-called path features from dependency trees. The path features collect POS tags or the first letter of POS tags along the tree between two nodes. Given two nodes $n_1$ and $n_2$, we denote the path feature as $path(n_1, n_2)$ and the coarse-grained path feature as $cpath(n_1, n_2)$. The syntactic head of a node $n$ is denoted as $n.h$.

We use the same feature templates for the online re-ordering and the two-stack-based systems, and they are slightly different from THMM. Figure 6 defines basic feature template functions. All feature templates are described here.

- **THMM system**: $f_{uni}(v_0), f_{uni}(c_1), g_{uni}(\beta_0), f_{context}(v_0), f_{context}(\beta_0), f_{pair-1}(v_0, \beta_0), f_{pair-1}(c_1, \beta_0), f_{pair}(v_0, c_1), f_{tri}(v_0, \beta_0, c_1), f_{tri-1}(v_0, \beta_0, c_0, lp), f_{tri-1}((v_0, \beta_0, c_0, rp), f_{tri-1}(c_0, v_0, c_0, lc), f_{tri-1}(c_0, \beta_0, c_0, lc), f_{tri-1}(c_0, \beta_0, \beta_0, lp), f_{tri-1}((c_0, \beta_0, \beta_0, lc), f_{tri-1}(c_1, \beta_0, c_1, lp), f_{tri-1}(c_1, \beta_0, \beta_0, lp), f_{tri-1}(c_1, \beta_0, c_1, lc), f_{tri-1}(c_1, \beta_0, c_0, lc), f_{tri-1}(c_0, \beta_0, c_0, lc), f_{quad}(v_0, \beta_0, v_0, rc), f_{quad}(c_0, \beta_0, v_0, lc), f_{quad}(c_0, \beta_0, c_0, lc), f_{quad}(c_0, \beta_0, c_0, rc), f_{quad}(c_0, \beta_0, c_0, lc2), f_{quad}(c_0, \beta_0, c_0, rc2), f_{quad}(c_0, \beta_0, \beta_0, lp, lc), f_{quad}(c_0, \beta_0, \beta_0, lc2), f_{quad}(c_1, c_0, \beta_0, c_1, rc), f_{quad}(c_1, c_0, \beta_0, c_1, lc2), f_{quad}(c_1, c_0, \beta_0, c_1, rc2).
Algorithm 5 Oracle generation for the two-stack–based system

1: procedure EXTRACTONEORACLE(c, A_{gold}, label)
2: \quad if c = (σ | i, σ_s, j | β, A) ∧ ¬∃k > j ∧ \exists[(i, l, k) ∈ A_{gold}] then
3: \quad \quad if label = nil then
4: \quad \quad \quad return REDUCE
5: \quad \quad else
6: \quad \quad \quad return ARC(label) ∘ REDUCE
7: \quad end if
8: \quad else if c = (σ | i, σ_s, j | β, A) ∧ ∃l[(i, l, j) ∈ A_{gold}] then
9: \quad \quad A_{gold} ← A_{gold} − (i, l, j)
10: \quad \quad return EXTRACTONEORACLE(c, σ_s, A_{gold}, label)
11: \quad else if c = (σ | i, σ_s, j | β, A) ∧ \exists i \forall i < l \land i ∈ σ \land \exists l[(i, l, j) ∈ A_{gold}] then
12: \quad \quad if label = nil then
13: \quad \quad \quad return MEM
14: \quad \quad else
15: \quad \quad \quad return ARC(label) ∘ MEM
16: \quad \quad end if
17: \quad else if c = (σ | i, σ_s, j | β, A) then
18: \quad \quad if label = nil then
19: \quad \quad \quad return RECALL
20: \quad \quad else
21: \quad \quad \quad return ARC(label) ∘ RECALL
22: \quad \quad end if
23: \quad end if
24: \quad if c = (σ, σ_s, j | β, A) then
25: \quad \quad if label = nil then
26: \quad \quad \quad return SHIFT
27: \quad \quad else
28: \quad \quad \quad return ARC(label) ∘ SHIFT
29: \quad \quad end if
30: \quad end if
31: \quad return nil
32: end procedure

f_{\text{quar}}(σ_1, β_0, β_0, lp, β_0, lc), f_{\text{quar}}(σ_1, β_0, β_0, lc, β_0, lc2), f_{\text{path}}(σ_0, β_0), f_{\text{path}}(σ_1, β_0), f_{\text{char}}(σ_0), f_{\text{char}}(β_0), f_{\text{char}}(β_1)

- Online re-ordering/two stack system: f_{\text{uni}}(σ_0), f_{\text{uni}}(σ_1), f_{\text{uni}}(σ_{0r}), g_{\text{uni}}(β_0), f_{\text{context}}(σ_0), f_{\text{context}}(β_0), f_{\text{pair}}(σ_0, β_0), f_{\text{pair}}(σ_1, β_0), f_{\text{pair}}(σ_{0r}, β_0), f_{\text{pair}}(σ_0, σ_1), f_{\text{pair}}(σ_0, σ_{0r}), f_{\text{tri}}(σ_0, β_0, σ_1), f_{\text{tri}}(σ_0, β_0, σ_{0r}), f_{\text{tri}}(σ_0, β_0, σ_1, lp), f_{\text{tri}}(σ_0, β_0, σ_{0r}, lp), f_{\text{tri}}(σ_0, β_0, σ_1, lc), f_{\text{tri}}(σ_0, β_0, σ_{0r}, lc), f_{\text{tri}}(σ_0, β_0, σ_1, lp, lc), f_{\text{tri}}(σ_0, β_0, σ_{0r}, lp, lc), f_{\text{tri}}(σ_0, β_0, σ_1, lp, lc), f_{\text{tri}}(σ_0, β_0, σ_{0r}, lp, lc), f_{\text{tri}}(σ_0, β_0, σ_1, rp), f_{\text{tri}}(σ_0, β_0, σ_{0r}, rp), f_{\text{tri}}(σ_0, β_0, σ_1, rp, lc), f_{\text{tri}}(σ_0, β_0, σ_{0r}, rp, lc), f_{\text{quar}}(σ_0, β_0, β_0, lp, σ_{0r}, σ_{0r}, rc), f_{\text{quar}}(σ_0, β_0, β_0, lp), f_{\text{quar}}(σ_0, β_0, β_0, lp, σ_0, lc), f_{\text{quar}}(σ_0, β_0, β_0, lp, σ_0, lc2), f_{\text{quar}}(σ_0, β_0, σ_0, rc, σ_0, rc2),
Computational Linguistics Volume 42, Number 3

Tree structures exhibit many computationally good properties, and parsing techniques for tree-structured representations are quite mature to some extent. When we consider semantics-oriented graphs, such as the representations for semantic role labeling.

Figure 6
Feature template functions.

4. Tree Approximation

Tree structures exhibit many computationally good properties, and parsing techniques for tree-structured representations are quite mature to some extent. When we consider semantics-oriented graphs, such as the representations for semantic role labeling.
Zhang, Du, Sun, and Wan  Transition-Based Parsing for Deep Dependency Structures

(SRL; Surdeanu et al. 2008; Hajič et al. 2009), CCG-grounded functor–argument (Clark, Hockenmaier, and Steedman 2002) analysis, HPSG-grounded predicate–argument analysis (Miyao, Ninomiya, and ichi Tsujii 2004), and reduction of MRS (Ivanova et al. 2012), syntactic trees can provide very useful features for semantic disambiguation (Punyakanok, Roth, and Yih 2008). Our parser also utilizes a path feature template (as defined in Section 3.3) to incorporate syntactic information for disambiguation.

In case syntactic tree information is not available, we introduce a tree approximation technique to induce tree backbones from deep dependency graphs. Such tree backbones can be utilized to train a tree parser which provides pseudo path features. In particular, we introduce an algorithm to associate every graph with a projective dependency tree, which we call weighted conversion. The tree reflects partial information about the corresponding graph. The key idea underlying this algorithm is to assign heuristic weights to all ordered pairs of words, and then find the tree with maximum weights. That means a tree frame of a given graph is automatically derived as an alternative for syntactic analysis.

We assign weights to all the possible edges (i.e., all pairs of words) and then determine which edges are to be kept by finding the maximum spanning tree. More formally, given a set of nodes \( V \), each possible edge \((i, j)\), where \( i, j \in V \), is assigned a heuristic weight \( \omega(i, j) \). Among all trees (denoted as \( T \)) over \( V \), the maximum spanning tree \( T_{\text{max}} \) contains the maximum sum of values of edges:

\[
T_{\text{max}} = \arg \max_{(V, A_T) \in T} \sum_{(i, j) \in A_T} t(i, j) \omega(i, j)
\]  

We separate the \( \omega(i, j) \) into three parts \( \omega(i, j) = A(i, j) + B(i, j) + C(i, j) \) that are as defined here.

- \( A(i, j) = a \cdot \max\{y(i, j), y(j, i)\} \): \( a \) is the weight for the existing edge on graph ignoring direction.
- \( B(i, j) = b \cdot y(i, j) \): \( b \) is the weight for the forward edge on the graph.
- \( C(i, j) = n - |i - j| \): This term estimates the importance of an edge where \( n \) is the length of the given sentence. For dependency parsing, we consider edges with short distance to be more important because those edges can be predicted more accurately in future parsing processes.
- \( a \gg b \gg n \) or \( a > bn > n^2 \): The converted tree should contain as many arcs as possible in original graph, and the direction of the arcs should not be changed if possible. The relationship of \( a, b, \) and \( c \) guarantees this.

After all edges are weighted, we can use maximum spanning tree algorithms to obtain the converted tree. To obtain the projective tree, we choose Eisner’s algorithm. For any graph, we can call this algorithm and get a corresponding tree. However, the tree is informative only when the given graph is dense enough. Fortunately, this condition holds for semantic dependency parsing.
5. Empirical Evaluation

5.1 Set-up

We present empirical evaluation of different incremental graph spanning algorithms for CCG-style functor–argument analysis, LFG-style grammatical relation analysis, and HPSG-style semantic dependency analysis for English and Chinese. Linguistically speaking, these types of syntacto-semantic dependencies directly encode information such as coordination, extraction, raising, control, as well as many other long-range dependencies. Experiments for a variety of formalisms and languages profile different aspects of transition-based deep dependency parsing models.

Figure 7 visualizes cross-format annotations assigned to the English sentence: A similar technique is almost impossible to apply to other crops, such as cotton, soybeans, and rice. This running example illustrates a range of linguistic phenomena such as coordination, verbal chains, argument and modifier prepositional phrases, complex noun phrases, and the so-called tough construction. The first format is from the popular corpus PropBank, which is widely used by various SRL systems. We can clearly see that compared with SRL, SDP uses dense graphs to represent much more syntacto-semantic information. This difference suggests to us that we should explore different algorithms for producing SRL and SDP graphs. Another thing worth noting is that, for the same phenomenon, annotation schemes may not agree with each other. Take the coordination construction, for example. For more details about the difference among different data sets, please refer to Ivanova et al. (2012).

For CCG analysis, we conduct experiments on English and Chinese CCGBank (Hockenmaier and Steedman 2007; Tse and Curran 2010). Following previous experimental set-up for English CCG parsing, we use Section 02-21 as training data, Section 00 as the development data, and Section 23 for testing. To conduct Chinese parsing experiments, we use data setting C of Tse and Curran (2012). For grammatical relation analysis, we conduct experiments on Chinese GRBank data (Sun et al. 2014). The selection for training, development, and test data is also according to Sun et al.’s (2014) experiments.

We also evaluate all parsing models using more HPSG-grounded semantics-oriented data, namely, DeepBank\textsuperscript{2} (Flickinger, Zhang, and Kordoni 2012) and EnjuBank (Miyao, Ninomiya, and ichi Tsujii 2004). Different from Penn Treebank–converted corpus, DeepBank’s annotations are essentially based on the parsing results given a large-scale linguistically precise HPSG grammar, namely, LingGO English resource grammar (ERG; Flickinger 2000), and manually disambiguated. As part of the full HPSG sign, the ERG also makes available a logical-form representation of propositional semantics, in the framework of minimal recursion semantics (MRS; Copestake et al. 2005). Such semantic information is reduced into variable-free bilexical dependency graphs (Oepen and Lønning 2006; Ivanova et al. 2012). In summary, DeepBank gives the reduction of logical-form meaning representations with respect to MRS. EnjuBank (Miyao, Ninomiya, and ichi Tsujii 2004) provides another corpus for semantic dependency parsing. This type of annotation is somehow shallower than DeepBank, given that only basic predicate–argument structures are concerned. Different from DeepBank but similar to CCGBank and GRBank, EnjuBank is semi-automatically converted from Penn Treebank–style annotations with linguistic heuristics. To conduct HPSG experiments, we use Sections 00 to 19 as training data and Section 20 as development data to tune parameters. For final

\textsuperscript{2} http://moin.delph-in.net/DeepBank.
A similar technique is almost impossible to apply to other crops, such as cotton, soybeans and rice.

(a) Format 1: Propositional semantics, from PropBank.

(b) Format 2: MRS-derived dependencies, from DeepBank HPSG annotations.

(c) Format 3: Predicate-argument structures, from Enju HPSG annotation.

(d) Format 4: Functor-argument structures, from CCGBank.

**Figure 7**
Dependency representations in (a) PropBank, (b) DeepBank, (c) Enju HPSGBank, and (d) CCGBank formats.

In the experiments, we use Sections 00 to 20 as training data and Section 21 as test data. The DeepBank and EnjuBank data sets are from SemEval 2014 Task 8 (Oepen et al. 2014), and the data splitting policy follows the shared task. Table 1 gives a summary of the data sets for experiments.

Experiments for English CCG-grounded analysis were performed using automatically assigned POS-tags that are generated by a symbol-refined generative HMM tagger\(^3\) (SR-HMM; Huang, Harper, and Petrov 2010). Experiments for English HPSG-grounded analysis used POS-tags provided by the shared task. For the experiments on Chinese CCGBank and GRBank, we use gold-standard POS tags.

We use the averaged perceptron algorithm with early update to estimate parameters, and beam search for decoding. We set the beam size to 16 and the number of iterations to 20 for all experiments. The measure for comparing two dependency graphs is precision and recall of tokens that are defined as \(\langle w_h, w_d, l \rangle\) tuples, where \(w_h\) is the head, \(w_d\) is the dependent, and \(l\) is the relation. Labeled precision/recall (LP/LR) is the ratio of tuples correctly identified by the automatic generator, and unlabeled precision/recall (UP/UR) is the ratio regardless of \(l\). F-score is a harmonic mean of precision and recall. These measures correspond to attachment scores (LAS/UAS) in dependency tree parsing and also used by the SemEval 2014 Task 8. The de facto standard

\(^3\) http://www.code.google.com/p/umd-featured-parser/.
Table 1
Data sets for experiments. Columns “Training” and “Test” present the number of sentences in training and test sets, respectively.

<table>
<thead>
<tr>
<th>Language</th>
<th>Formalism</th>
<th>Data</th>
<th>Training</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>CCG</td>
<td>CCGBank</td>
<td>39,604</td>
<td>2,407</td>
</tr>
<tr>
<td></td>
<td>HPSG</td>
<td>DeepBank</td>
<td>34,003</td>
<td>1,348</td>
</tr>
<tr>
<td></td>
<td>HPSG</td>
<td>EnjuBank</td>
<td>34,003</td>
<td>1,348</td>
</tr>
<tr>
<td>Chinese</td>
<td>CCG</td>
<td>CCGBank</td>
<td>22,339</td>
<td>2,813</td>
</tr>
<tr>
<td></td>
<td>LFG</td>
<td>GRBank</td>
<td>22,277</td>
<td>2,557</td>
</tr>
</tbody>
</table>

to evaluate CCG parsers also considers supertags. Because no supertagging is performed in our experiments, only the unlabeled precision/recall/F-score is comparable to the results reported in other papers. And the labeled performance reported here only considers the labels assigned to dependency arcs that indicate the argument types. For example, an arc label \( \text{arg1} \) denotes that the dependent is the first argument of the head.

5.2 Parsing Efficiency

We evaluate the real running time of our final trained parser using realistic data. The test sentences are collected from English Wikipedia and Chinese Gigaword (LDC2005T14). First, we show the influence of beam size in Figure 8. In this experiment, the DeepBank trained models are used for test. We can see that the parsers run in nearly linear time regardless of the beam width in realistic situations. Second, we report the the averaged real running time of models trained on different data sets in Figure 9. Again, we can see that the parser runs in close to linear time for a variety of linguistically motivated representations. The results also suggest that our proposed transition-based parsers can automatically learn the complexity of linguistically motivated dependency structures from an annotated corpus. Note that although within the deep parsing framework, the study of formal grammars is partially relevant for data-driven dependency parsing,

![Figure 8](http://example.com/figure8.png)

**Figure 8**
Real running time relative to beam size. Tested using DeepBank-trained models.
where our parsers rely on inductive inference from treebank data, and only *implicitly* use a grammar.

### 5.3 Importance of Transition Combination

Figure 10 and Table 2 summarize the labeled parsing results on all of the five data sets. In this experiment, we distinguish parsing models with and without transition combination. All models take only the surface word form and POS tag information and do not derive features from any syntactic analysis. The importance of transition combination is highlighted by the comparative evaluation on parsers using this mechanism or not. Significant improvements are observed over a wide range of conditions: Parsers based on different transition systems for different languages and different formalisms almost always benefit. This result suggests a necessary strategy for designing transition systems for producing deep dependency graphs: Configurations should be essentially modified by every transition.

Because of the importance of transition combination, all the following experiments utilize the transition combination strategy.

### 5.4 Model Diversity and Parser Ensemble

#### 5.4.1 Model Diversity.

For model ensemble, besides the accuracy of each single model, it is also essential that the models to be integrated should be very different. We argue that heterogeneous parsing models can be built by varying the underlying transition systems. By reversing the sentence from right to left, we can build other model variants with the same transition system. To evaluate the differences between two models $A$ and $B$, we define the following metric:

$$\frac{2 \times |D_A \cap D_B|}{|D_A| + |D_B|}$$

where $D_X$ denotes the set of dependencies related to held out sentences returned by model $X$. Tables 3 and 4 show the model diversity evaluated on English and Chinese data, respectively. We can see that parsing models built upon different transition systems do vary. Even for one specific transition system, different processing directions yield quite different parsing results.
Figure 10
Labeled parsing F-scores of different transition system with and without transition combination. “Standard” denotes the standard systems, which do not combine an ARC transition with its following transition.

5.4.2 Parser Ensemble. Parser ensemble has been shown very effective to boost the performance of data-driven tree parsers (Nivre and McDonald 2008; Surdeanu and Manning 2010; Sun and Wan 2013). Empirically, the two proposed systems together with the existing THMM system exhibit complementary prediction powers, and their combination yields superior accuracy. We present a simple yet effective voting strategy for parser ensemble. For each pair of words in each sentence, we count the number of models that give positive predictions. If the number is greater than a threshold (we set it to half the number of models in this work), we put this arc to the final graph, and label the arc with the most common label of what the models give.

Table 5 presents the parsing accuracy of the combined model where six base models are utilized for voting. We can see that a system ensemble is quite helpful. Given that our graph parsers all run in expected linear time, the combined system also runs very efficiently.

5.5 Impact of Syntactic Parsing
5.5.1 Effectiveness of Syntactic Features. Syntactic parsing, especially the full one, has been shown very important for boosting the performance of SRL, a well studied shallow semantic parsing task (Punyakanok, Roth, and Yih 2008). According to the comprehensive evaluation presented in Punyakanok, Roth, and Yih (2008) and Zhuang and Zong
Table 2
Performance of different transition system with and without transition combination on the test set of the DeepBank/EnjuBank data, on the development set of the English and Chinese CCGBank data, and on the development set of the Chinese GRBank data. $S_{\text{std}}$ denotes the standard system, which does not combine an ARC transition with its following transition.

<table>
<thead>
<tr>
<th></th>
<th>DeepBank</th>
<th>EnjuBank</th>
<th>CCGBank</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LP</td>
<td>LR</td>
<td>LF</td>
</tr>
<tr>
<td>$S_{\text{std}}$</td>
<td>82.71%</td>
<td>84.32%</td>
<td>83.51%</td>
</tr>
<tr>
<td>$S_T$</td>
<td>85.00%</td>
<td>84.40%</td>
<td>84.70%</td>
</tr>
<tr>
<td>$S_{\text{std}}$</td>
<td>82.60%</td>
<td>84.46%</td>
<td>83.52%</td>
</tr>
<tr>
<td>$S_S$</td>
<td>84.63%</td>
<td>83.98%</td>
<td>84.30%</td>
</tr>
<tr>
<td>$S_{\text{std}}$</td>
<td>82.97%</td>
<td>84.65%</td>
<td>83.80%</td>
</tr>
<tr>
<td>$S_{2S}$</td>
<td>85.01%</td>
<td>84.41%</td>
<td>84.71%</td>
</tr>
</tbody>
</table>

For the English DeepBank and EnjuBank data, the performance of the standard system ($S_{\text{std}}$) is compared to the performance of the combined transition system ($S_T$). The performance is measured in terms of LP, LR, and LF, which are percentages of correct tokens. The table shows that the combined transition system outperforms the standard system on all metrics.

For the Chinese CCGBank and GRBank data, the performance is measured in a similar manner. The table shows that the combined transition system also outperforms the standard system on all metrics.

When we consider semantics-oriented deep dependency structures, including the representations for CCG-grounded functor–argument (Clark, Hockenmaier, and Steedman 2002) analysis, HPSG-grounded predicate–argument analysis (Miyao, Ninomiya, and ichi Tsujii 2004), and reduction of MRS (Ivanova et al. 2012), syntactic parses can also provide very useful features for disambiguation. To evaluate the impact of syntactic tree parsing, we include more features, namely, path features, to our parsing models. The detailed description of syntactic features are presented in Section 3.3. In this work, we apply syntactic dependency parsers rather than phrase-structure parsers. Figure 11 summarizes the impact of features derived from syntactic trees. We can clearly see that syntactic features are effective to enhance semantic dependency parsing. These informative features lead to on average 1.14% and 1.03% absolute improvements for English and Chinese CCG parsing. Compared with SRL, the improvement brought by syntactic parsing is smaller. We think one main reason for this difference is the information density of different types of graphs. SRL graphs usually annotate only on verbal predicates and their nominalization, whereas the semantic graphs grounded by CCG and HPSG target all words. In other words, SRL provides partial analysis and semantic dependency parsing provides full analysis. Accordingly, SRL needs structural information generated by a syntactic parser much more than semantic dependency parsing.
Table 3
Model diversity between different models on the test set of the DeepBank/EnjuBank data and on the development set of the English CCGBank data. $S_\text{rev}$ means processing a sentence with system $S_\text{s}$ but in the right-to-left word order.

<table>
<thead>
<tr>
<th></th>
<th>DeepBank</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_T$</td>
<td>0.9285</td>
<td>0.9285</td>
<td>0.8788</td>
<td>0.8796</td>
<td>0.8797</td>
</tr>
<tr>
<td>$S_S$</td>
<td>0.9385</td>
<td>0.8748</td>
<td>0.8776</td>
<td>0.8773</td>
<td></td>
</tr>
<tr>
<td>$S_2$</td>
<td>0.8772</td>
<td>0.8802</td>
<td>0.8790</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$S_\text{rev}$</td>
<td>0.9390</td>
<td>0.9364</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$S_\text{rev}$</td>
<td></td>
<td></td>
<td>0.9413</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>EnjuBank</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_T$</td>
<td>0.9504</td>
<td>0.9481</td>
<td>0.9045</td>
<td>0.9038</td>
<td>0.9043</td>
</tr>
<tr>
<td>$S_S$</td>
<td>0.9503</td>
<td>0.9046</td>
<td>0.9060</td>
<td>0.9055</td>
<td></td>
</tr>
<tr>
<td>$S_2$</td>
<td>0.9066</td>
<td>0.9087</td>
<td>0.9076</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$S_\text{rev}$</td>
<td>0.9562</td>
<td>0.9565</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$S_\text{rev}$</td>
<td></td>
<td></td>
<td>0.9584</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>CCGBank</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_T$</td>
<td>0.9547</td>
<td>0.9532</td>
<td>0.9155</td>
<td>0.9164</td>
<td>0.9182</td>
</tr>
<tr>
<td>$S_S$</td>
<td>0.9575</td>
<td>0.9166</td>
<td>0.9179</td>
<td>0.9187</td>
<td></td>
</tr>
<tr>
<td>$S_2$</td>
<td>0.9200</td>
<td>0.9197</td>
<td>0.9205</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$S_\text{rev}$</td>
<td>0.9586</td>
<td>0.9575</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$S_\text{rev}$</td>
<td></td>
<td></td>
<td>0.9617</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4
Model diversity between different models on the development set of the Chinese CCGBank/GRBank data.

<table>
<thead>
<tr>
<th></th>
<th>CCGBank</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_T$</td>
<td>0.9261</td>
<td>0.9262</td>
<td>0.8668</td>
<td>0.8614</td>
<td>0.8658</td>
</tr>
<tr>
<td>$S_S$</td>
<td>0.9314</td>
<td>0.8667</td>
<td>0.8593</td>
<td>0.8663</td>
<td></td>
</tr>
<tr>
<td>$S_2$</td>
<td>0.8694</td>
<td>0.8624</td>
<td>0.8683</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$S_\text{rev}$</td>
<td>0.9130</td>
<td>0.9107</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$S_\text{rev}$</td>
<td></td>
<td></td>
<td>0.9230</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>GRBank</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_T$</td>
<td>0.8918</td>
<td>0.8861</td>
<td>0.8398</td>
<td>0.8301</td>
<td>0.8328</td>
</tr>
<tr>
<td>$S_S$</td>
<td>0.9058</td>
<td>0.8455</td>
<td>0.8378</td>
<td>0.8414</td>
<td></td>
</tr>
<tr>
<td>$S_2$</td>
<td>0.8460</td>
<td>0.8391</td>
<td>0.8441</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$S_\text{rev}$</td>
<td>0.8969</td>
<td>0.8984</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$S_\text{rev}$</td>
<td></td>
<td></td>
<td>0.9111</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 5
Performance of base and combined models on the test set of the DeepBank/EnjuBank data, on the development set of the English and Chinese CCGBank data and on the development set of the Chinese GRBank data. Note that the labeled results for CCG parsing do not consider supertags.

<table>
<thead>
<tr>
<th></th>
<th>English</th>
<th></th>
<th></th>
<th>Chinese</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DeepBank</td>
<td>UP</td>
<td>UR</td>
<td>UF</td>
<td>LP</td>
<td>LR</td>
</tr>
<tr>
<td>ST</td>
<td>87.03%</td>
<td>86.42%</td>
<td>86.72</td>
<td>85.00%</td>
<td>84.40%</td>
<td>84.70</td>
</tr>
<tr>
<td>ST\textsuperscript{rev}</td>
<td>88.16%</td>
<td>88.16%</td>
<td>88.16</td>
<td>86.12%</td>
<td>86.11%</td>
<td>86.12</td>
</tr>
<tr>
<td>S\textsubscript{S}</td>
<td>86.57%</td>
<td>85.91%</td>
<td>86.24</td>
<td>84.63%</td>
<td>83.98%</td>
<td>84.30</td>
</tr>
<tr>
<td>S\textsubscript{S}\textsuperscript{rev}</td>
<td>88.53%</td>
<td>88.27%</td>
<td>88.40</td>
<td>86.63%</td>
<td>86.38%</td>
<td>86.51</td>
</tr>
<tr>
<td>S\textsubscript{2S}</td>
<td>86.99%</td>
<td>86.38%</td>
<td>86.68</td>
<td>85.01%</td>
<td>84.41%</td>
<td>84.71</td>
</tr>
<tr>
<td>S\textsubscript{2S}\textsuperscript{rev}</td>
<td>88.12%</td>
<td>88.11%</td>
<td>88.11</td>
<td>86.28%</td>
<td>86.26%</td>
<td>86.27</td>
</tr>
<tr>
<td>Combined</td>
<td>88.29%</td>
<td>90.27%</td>
<td>89.27</td>
<td>86.46%</td>
<td>88.40%</td>
<td>87.42</td>
</tr>
</tbody>
</table>
Table 6
Performance of English and Chinese SRL achieved by representative full and shallow parsing-based systems. The results are copied from Punyakanok, Roth, and Yih (2008) and Zhuang and Zong (2010).

<table>
<thead>
<tr>
<th></th>
<th>Precison</th>
<th>Recall</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Full parsing</td>
<td>77.09%</td>
<td>75.51%</td>
<td>76.29%</td>
</tr>
<tr>
<td>Shallow parsing</td>
<td>75.48%</td>
<td>67.13%</td>
<td>71.06%</td>
</tr>
<tr>
<td>Chinese</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Full parsing</td>
<td>79.17%</td>
<td>72.09%</td>
<td>75.47%</td>
</tr>
<tr>
<td>Shallow parsing</td>
<td>72.57%</td>
<td>67.02%</td>
<td>69.68%</td>
</tr>
</tbody>
</table>

5.5.2 Comparison of Different Tree Parsers. There are two dominant data-driven approaches to syntactic dependency tree parsing: transition-based (Yamada and Matsumoto 2003; Nivre 2008) and graph-based (McDonald 2006; Torres Martins, Smith, and Xing 2009). In terms of overall per token prediction, the transition-based and graph-based tree parsers achieve comparable performance (Suzuki et al. 2009; Weiss et al. 2015). To evaluate the impact of the two tree parsing approaches on semantic dependency parsing, we use two tree parsers to serve our graph parser. The first one is our in-house implementation of the algorithm presented in Zhang and Nivre (2011), and the second one is a second-order graph-based parser (Bohnet 2010). The tree parsers are trained with the unlabeled tree annotations provided by the English and Chinese CCGBank data. For both English and Chinese experiments, 5-fold cross validation is performed to parse the training data to avoid overfitting. The accuracy of tree parsers is shown in Table 7. Results presented in Figure 12 indicate that the two parsers are also equivalently effective for producing semantic analysis. This result is somehow non-obvious given that the combination of a graph-based and transition-based parser usually gives significantly better parsing performance (Nivre and McDonald 2008; Torres Martins et al. 2008).

5.6 Effectiveness of Tree Approximation

In case syntactic information is not available, we propose a tree approximation technique to induce tree backbones from deep dependency graphs. In particular, our technique guarantees that the automatically derived trees are projective, which is a necessary condition for a number of effective tree parsing algorithms. We can utilize these pseudo trees as an alternative to syntactic analysis. To evaluate the effectiveness of tree approximation, we compare the contribution to semantic dependency parsing of syntactic trees and pseudo trees. In this experiment, we use a transition-based tree parser to generate automatic analysis. Figure 13 presents the results. Generally speaking, pseudo trees contribute to semantic dependency parsing equally well as syntactic trees. Sometimes, they perform even better. There is a considerable drop when DeepBank data are applied. We think the main reason is the density of DeepBank graphs. Because there are fewer edges in the original graphs, it is harder to extract informative pseudo trees. As a result, the final graph parsing benefits less.

4 http://www.code.google.com/p/mate-tools/
Figure 11
Parsing accuracy with and without syntactic features. The syntactic trees for experiments on
DeepBank and EnjuBank data sets are provided by the SemEval 2014 shared task, and they are
automatically generated by the Stanford Parser. The syntactic trees for experiments on English
and Chinese CCG data sets are generated by our in-house implementation of the model
introduced in Zhang and Nivre (2011).

It is also possible to build a parser ensemble on pseudo tree enhanced models. However, the effectiveness of system combination is not as effective as integrating non-tree models. Table 8 summarizes the detailed parsing accuracy. We can see that system ensemble is still helpful, though the improvement is limited.
Table 7
Accuracy of preprocessing on the development data for CCG analysis. Tr and Gr, respectively, denote transition-based and graph-based tree parsers.

<table>
<thead>
<tr>
<th></th>
<th>UAS(Tr)</th>
<th>UAS(Gr)</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>93.48%</td>
<td>93.47%</td>
</tr>
<tr>
<td>Chinese</td>
<td>80.97%</td>
<td>80.81%</td>
</tr>
</tbody>
</table>

5.7 Comparison with Other Parsers
5.7.1 Comparison with Grammar-Based Parsers. We compare our parser with several representative Treebank-guided, grammar-based parsers that achieve state-of-the-art performance for CCG and HPSG analysis. The grammar-based parsers selected represent two different architectures.

- The first type of parser implements a shift-reduce parsing architecture and also uses beam search for practical decoding. In particular, we compare our parser with the state-of-the-art CCG parser introduced in Xu, Clark, and

![Figure 12](http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00252)
Zhang (2014). This parser extends a shift-reduce CFG parser (Zhang and Clark 2011a) with a dependency model.

- The second type of parser implements the chart parsing architecture with some refinements. For CCG analysis, we focus on the parser proposed by

---

5 The unlabeled parsing results are not reported in the original paper. The figures presented in Table 9 are provided by Wenduan Xu.
Table 8
Performance of base and combined models on the test set of the DeepBank/EnjuBank data and on the development set of the English and Chinese CCGBank data. Features extracted from \textit{pseudo} trees are utilized for disambiguation.

<table>
<thead>
<tr>
<th></th>
<th>DeepBank</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>UP</td>
<td>UR</td>
<td>UF</td>
<td>LP</td>
<td>LR</td>
<td>LF</td>
<td>UP</td>
<td>UR</td>
<td>UF</td>
<td>LP</td>
<td>LR</td>
<td>LF</td>
<td>UP</td>
<td>UR</td>
<td>UF</td>
</tr>
<tr>
<td>English</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DeepBank</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S_T</td>
<td>87.99%</td>
<td>87.64%</td>
<td>87.81%</td>
<td>85.95%</td>
<td>85.61%</td>
<td>85.78%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S_{TP}</td>
<td>88.94%</td>
<td>88.98%</td>
<td>88.96%</td>
<td>87.07%</td>
<td>87.11%</td>
<td>87.09%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S_S</td>
<td>87.76%</td>
<td>87.45%</td>
<td>87.60%</td>
<td>85.83%</td>
<td>85.52%</td>
<td>85.67%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S_{SP}</td>
<td>88.72%</td>
<td>88.65%</td>
<td>88.69%</td>
<td>86.88%</td>
<td>86.82%</td>
<td>86.85%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S_{SS}</td>
<td>87.92%</td>
<td>87.60%</td>
<td>87.76%</td>
<td>86.03%</td>
<td>85.72%</td>
<td>85.87%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S_{2S}</td>
<td>89.04%</td>
<td>88.85%</td>
<td>88.95%</td>
<td>87.15%</td>
<td>86.96%</td>
<td>87.05%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Combined</td>
<td>88.54%</td>
<td>90.25%</td>
<td>89.39%</td>
<td>86.65%</td>
<td>88.32%</td>
<td>87.48%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EnjuBank</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S_T</td>
<td>91.88%</td>
<td>91.45%</td>
<td>91.66%</td>
<td>90.60%</td>
<td>90.17%</td>
<td>90.38%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S_{TP}</td>
<td>92.82%</td>
<td>92.83%</td>
<td>92.83%</td>
<td>91.61%</td>
<td>91.61%</td>
<td>91.61%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S_S</td>
<td>91.76%</td>
<td>91.39%</td>
<td>91.58%</td>
<td>90.50%</td>
<td>90.14%</td>
<td>90.32%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S_{SP}</td>
<td>92.66%</td>
<td>92.65%</td>
<td>92.65%</td>
<td>91.45%</td>
<td>91.44%</td>
<td>91.45%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S_{SS}</td>
<td>91.85%</td>
<td>91.54%</td>
<td>91.70%</td>
<td>90.63%</td>
<td>90.33%</td>
<td>90.48%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S_{2S}</td>
<td>92.92%</td>
<td>92.83%</td>
<td>92.87%</td>
<td>91.77%</td>
<td>91.68%</td>
<td>91.73%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Combined</td>
<td>92.47%</td>
<td>93.52%</td>
<td>92.99%</td>
<td>91.34%</td>
<td>92.38%</td>
<td>91.86%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CCGBank</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S_T</td>
<td>92.15%</td>
<td>91.05%</td>
<td>91.60%</td>
<td>88.20%</td>
<td>87.15%</td>
<td>87.67%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S_{TP}</td>
<td>92.46%</td>
<td>92.27%</td>
<td>92.37%</td>
<td>88.78%</td>
<td>88.61%</td>
<td>88.70%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S_S</td>
<td>91.91%</td>
<td>91.18%</td>
<td>91.64%</td>
<td>87.97%</td>
<td>87.28%</td>
<td>87.62%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S_{SP}</td>
<td>92.34%</td>
<td>92.43%</td>
<td>92.39%</td>
<td>88.67%</td>
<td>88.76%</td>
<td>88.72%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S_{SS}</td>
<td>91.86%</td>
<td>91.13%</td>
<td>91.49%</td>
<td>87.92%</td>
<td>87.22%</td>
<td>87.57%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S_{2S}</td>
<td>92.53%</td>
<td>92.41%</td>
<td>92.47%</td>
<td>88.85%</td>
<td>88.73%</td>
<td>88.79%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Combined</td>
<td>92.38%</td>
<td>93.20%</td>
<td>92.79%</td>
<td>88.92%</td>
<td>89.71%</td>
<td>89.31%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

|                |        |        |        |        |        |        |        |        |        |        |        |        |        |        |        |
| English        |        |        |        |        |        |        |        |        |        |        |        |        |        |        |        |
| DeepBank       |        |        |        |        |        |        |        |        |        |        |        |        |        |        |        |
| S_T            | 87.44% | 86.41% | 86.93% | 83.44% | 82.43% | 82.94% |         |        |        |        |        |        |         |        |        |
| S_{TP}         | 87.11% | 87.04% | 87.07% | 83.24% | 83.17% | 83.21% |         |        |        |        |        |        |         |        |        |
| S_S            | 86.76% | 86.52% | 86.64% | 82.85% | 82.63% | 82.74% |         |        |        |        |        |        |         |        |        |
| S_{SP}         | 86.46% | 87.54% | 87.00% | 82.69% | 83.72% | 83.20% |         |        |        |        |        |        |         |        |        |
| S_{SS}         | 87.09% | 86.91% | 87.00% | 83.21% | 83.03% | 83.12% |         |        |        |        |        |        |         |        |        |
| S_{2S}         | 86.57% | 87.69% | 87.13% | 82.75% | 83.82% | 83.28% |         |        |        |        |        |        |         |        |        |
| Combined       | 87.27% | 89.00% | 88.12% | 83.57% | 85.23% | 84.39% |         |        |        |        |        |        |         |        |        |

Auli and Lopez (2011b). The basic system architecture follows the well-engineered C&C Parser,\(^6\) and additionally applies a number of advanced machine learning and optimization techniques, including belief propagation, dual decomposition Auli and Lopez (2011a), and parameter estimation with softmax-margin loss (Auli and Lopez 2011b), to enhance the results. For HPSG analysis, we compare with the well-studied Enju

---

\(^6\) [http://svn.ask.it.usyd.edu.au/trac/candc](http://svn.ask.it.usyd.edu.au/trac/candc)
Table 9

Parsing results on test sets obtained by representative parsers. State-of-the-art results on these data sets, as reported in Oepen et al. (2014), Martins and Almeida (2014), Xu, Clark, and Zhang (2014), Auli and Lopez (2011b), Du, Sun, and Wan (2015), Sun et al. (2014), are included.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>LP</th>
<th>LR</th>
<th>LF</th>
</tr>
</thead>
<tbody>
<tr>
<td>DeepBank</td>
<td>S*rev</td>
<td>86.28%</td>
<td>86.26%</td>
</tr>
<tr>
<td></td>
<td>Combined</td>
<td>86.46%</td>
<td>88.40%</td>
</tr>
<tr>
<td></td>
<td>S*rev+Pseudo Tree</td>
<td>87.15%</td>
<td>86.96%</td>
</tr>
<tr>
<td></td>
<td>Combined+Pseudo Tree</td>
<td>86.65%</td>
<td>88.32%</td>
</tr>
<tr>
<td>Factorization (Turbo)</td>
<td>(Martins and Almeida 2014)</td>
<td>88.82%</td>
<td>87.35%</td>
</tr>
<tr>
<td>EnjuBank</td>
<td>S*rev</td>
<td>90.88%</td>
<td>90.80%</td>
</tr>
<tr>
<td></td>
<td>Combined</td>
<td>90.15%</td>
<td>92.43%</td>
</tr>
<tr>
<td></td>
<td>S*rev+Pseudo Tree</td>
<td>91.77%</td>
<td>91.68%</td>
</tr>
<tr>
<td></td>
<td>Combined+Pseudo Tree</td>
<td>91.34%</td>
<td>92.38%</td>
</tr>
<tr>
<td>Factorization (Turbo)</td>
<td>(Martins and Almeida 2014)</td>
<td>91.95%</td>
<td>89.92%</td>
</tr>
<tr>
<td>English CCGBank</td>
<td>S*rev</td>
<td>91.84%</td>
<td>91.75%</td>
</tr>
<tr>
<td></td>
<td>Combined</td>
<td>92.06%</td>
<td>93.14%</td>
</tr>
<tr>
<td></td>
<td>S*rev+Pseudo Tree</td>
<td>92.49%</td>
<td>92.30%</td>
</tr>
<tr>
<td></td>
<td>Combined+Pseudo Tree</td>
<td>92.52%</td>
<td>93.13%</td>
</tr>
<tr>
<td>Factorization (Turbo)</td>
<td>(Martins and Almeida 2014)</td>
<td>93.15%</td>
<td>91.06%</td>
</tr>
<tr>
<td>Chinese GRBank</td>
<td>S*rev</td>
<td>82.28%</td>
<td>83.11%</td>
</tr>
<tr>
<td></td>
<td>Combined</td>
<td>84.92%</td>
<td>85.28%</td>
</tr>
<tr>
<td>Transition-based</td>
<td>(Sun et al. 2014)</td>
<td>83.93%</td>
<td>79.82%</td>
</tr>
<tr>
<td>Chinese CCGBank</td>
<td>S*rev</td>
<td>85.07%</td>
<td>86.02%</td>
</tr>
<tr>
<td></td>
<td>Combined</td>
<td>86.35%</td>
<td>88.85%</td>
</tr>
<tr>
<td></td>
<td>S*rev+Pseudo Tree</td>
<td>86.65%</td>
<td>87.34%</td>
</tr>
<tr>
<td></td>
<td>Combined+Pseudo Tree</td>
<td>87.14%</td>
<td>88.60%</td>
</tr>
</tbody>
</table>

Parser,\(^7\) which develops a number of advanced techniques for discriminative deep parsing—for example, maximum entropy estimation with feature forest (Miyao and Tsujii 2008) and efficient decoding with supertagging and CFG-filtering (Matsuzaki, Miyao, and Tsujii 2007).

Table 9 shows the final results on the test data for each data set. The representative shift-reduce parser for comparison utilizes a very similar learning and decoding architectures to our system. Similar to our parser, Xu, Clark, and Zhang’s (2014) parser incrementally processes a sentence and uses a beam decoder that performs an inexact

---

search. Xu, Clark, and Zhang’s parser sets beam width to 128, while ours is 16. It also uses the structured prediction algorithm for parameter estimation. The major difference is that the shift-reduce CCG parser explicitly utilizes a core grammar to guide decoding, whereas our parser excludes all such information. Actually, our models reported here also exclude all syntactic information because no syntactic parse is used for feature extraction. We can see that our individual system based on the two stack transition system achieves equivalent performance to the CCG-driven parser. Moreover, when this individual system is augmented with tree approximation, the accuracy is significantly improved. Note that the individual system with both settings does not rely on any explicit syntactic information. This result on one hand indicates the effectiveness of adapting syntactic parsing techniques for full semantic parsing, and on the other hand suggests the possibility of using semantically structural (not syntactically structural) information only to achieve high-accuracy semantic parsing.

Statistical parsers based on chart parsing are able to perform a more principled search and therefore usually achieve better parsing accuracy than a normal shift-reduce parser. We also compare our parsing models with two state-of-the-art chart parsers, namely, the Enju Parser (Miyao and Tsujii 2008) and Auli and Lopez’s (2011b) parser. Different from Xu, Clark, and Zhang’s (2014) shift-reduce parser and our models, Auli and Lopez’s (2011b) parser does not guarantee to produce analysis for arbitrary sentences. Usually, the numerical performance evaluated on all sentences is lower than the results obtained on sentences that can be parsed. Note that Auli and Lopez (2011b) only reported results on sentences that are covered, whereas Oepen et al. (2014) reported results on all sentences, which is achieved by Enju Parser. From Table 9, we can clearly see that our graph-spanning models are very competitive. The best individual and combined models outperform the Enju Parser and perform equally well to Auli and Lopez’s (2011b) parser. It is worth noting that strictly less information is used by our parsers.

5.7.2 Comparison with Other Data-Driven Parsers. We also compare our parser with recently developed data-driven, factorization models (Martins and Almeida 2014; Du, Sun, and Wan 2015). Different from projective but similar to non-projective tree parsing, decoding for factorization models where very basic second-order sibling factors are incorporated is NP-hard. See the proof presented in our early work (Du, Sun, and Wan 2015) for details. To perform principled decoding, dual decomposition is used and achieves good empirical results (Martins and Almeida 2014; Du, Sun, and Wan 2015).

From Table 9, we can see that the transition-based approach augmented with tree approximation is comparable to the factorization approach in general. Compared with the Turbo Parser, our individual and hybrid models perform significantly worse on DeepBank but significantly better on EnjuBank. We think one main reason is because of the annotation styles. Though both corpora are based on HPSG, the annotations in question are quite different. DeepBank graphs are more sparse than EnjuBank, which makes tree approximation less effective. It seems that the transition-based parser suffers more when fewer output edges are targeted. The two approaches achieve equivalent performance for CCG parsing.

6. Related Work

Deep linguistic processing is concerned with NLP approaches that aim at modeling the complexity of natural languages in rich linguistic representations. Such approaches are typically related to a particular computational linguistic theory (e.g., CCG, LFG, and
HPSG). Parsing in these formalisms provides an elegant way to generate deep syntacto-semantic dependency structures with high quality (Clark and Curran 2007; Miyao, Sagae, and Tsujii 2007; Miyao and Tsujii 2008). The incremental shift-reduce parsing architecture has been implemented for CCG parsing (Zhang and Clark 2011a; Ambati et al. 2015). Besides using phrase-structure rules only, a shift-reduce parser can be enhanced by incorporating a dependency model (Xu, Clark, and Zhang 2014). Our parser and the two above parsers have some essential resemblances, including learning and decoding algorithms. The main difference is the usage of syntactic and grammatical information. The comparison in Section 5.7 gives a rough idea of the impact of explicitly using grammatical constraints. A deep-grammar-guided parsing model usually cannot produce full coverage and the time complexity of the corresponding parsing algorithms is very high. Some NLP applications may favor lightweight solutions to build deep dependency structures.

Different from grammar-guided approaches, data-driven approaches make essential use of machine learning from linguistic annotations in order to parse new sentences. Such approaches, for example, transition-based (Yamada and Matsumoto 2003; Nivre 2008) and graph-based (McDonald 2006; Torres Martins, Smith, and Xing 2009) models, have attracted the most attention of dependency parsing in recent years. Several successful parsers (e.g., MST, Mate, and Malt parsers) have been built and applied to many NLP applications. Recently, two advanced techniques have been studied to enhance a transition-based parser. First, developing features has been shown crucial to advancing parsing accuracy and a very rich feature set is carefully evaluated by Zhang and Nivre (2011). Second, beyond deterministic greedy search, beam search and principled dynamic programming strategies have been used to explore more possible hypotheses (Zhang and Clark 2008; Huang and Sagae 2010). When we implement our graph parser, we also leverage rich features and beam search to obtain good parsing accuracy.

Most research concentrated on surface dependency structures, and the majority of existing approaches are limited to producing only tree-shaped graphs. We notice three distinguished exceptions in early work. Sagae and Tsujii (2008) proposed a DAG parser that is able to handle projective directed dependency graphs, and that uses the pseudo-projective parsing technique (Nivre and Nilsson 2005) to build crossing arcs. Titov et al. (2009) and Henderson et al. (2013) introduced non-planar parsing to parse PropBank (Palmer, Gildea, and Kingsbury 2005) structures. However, neither technique handles crossing arcs fully well. There have been a number of papers trying to build non-projective trees, which inspired the design of our transition systems. Especially, we borrow key ideas from Nivre (2009), Gómez-Rodríguez and Nivre (2010), and Gómez-Rodríguez and Nivre (2013). In addition to the investigation on the transition-based approach, McDonald and Pereira (2006) presented a factorization parser that can generate dependency graphs in which a word may depend on multiple heads, and evaluated it on the Danish Treebank. Very recently, the dual decomposition technique has been adopted to achieve principled decoding for factorization models. High-accuracy models have been introduced in Martins and Almeida (2014) and Du, Sun, and Wan (2015).

7. Conclusion

We study transition-based approaches that produce general dependency graphs directly from input sequences of words, in a way nearly as simple as tree parsers. We introduce two new graph-spanning algorithms to generate arbitrary directed graphs, which suit deep dependency parsing well. We also introduce transition combination and tree
approximation for statistical disambiguation. Statistical parsers built upon these new techniques have been evaluated with dependency structures that are extracted from linguistically deep CCG, LFG, and HPSG derivations. Our models achieve state-of-the-art performance on five representative data sets for English and Chinese parsing. Experiments demonstrate the effectiveness of grammar-free, transition-based approaches to dealing with complex linguistic phenomena beyond surface syntax.

In addition to deep dependency parsing, many other NLP tasks (e.g., quantifier scope disambiguation [Manshadi, Gildea, and Allen 2013] and event extraction [Li, Ji, and Huang 2013]), can be formulated as graph spanning problems. We think such tasks can benefit from algorithms that span general graphs rather than trees, and our new transition-based parsers can provide practical solutions to these tasks.

Acknowledgments
This work was supported by the National Natural Science Foundation of China under grants 61300064 and 61331011, and the National High-Tech R&D Program under grant 2015AA015403. We are very grateful to the anonymous reviewers for their insightful and constructive comments and suggestions.

References


Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 1545–1555, Beijing.


McDonald, Ryan and Fernando Pereira. 2006. Online learning of approximate dependency parsing algorithms. In


