LSTM Shift-Reduce CCG Parsing

Wenduan Xu
Computer Laboratory
University of Cambridge
wx217@cam.ac.uk

Abstract
We describe a neural shift-reduce parsing model for CCG, factored into four unidirectional LSTMs and one bidirectional LSTM. This factorization allows the linearization of the complete parsing history, and results in a highly accurate greedy parser that outperforms all previous beam-search shift-reduce parsers for CCG. By further deriving a globally optimized model using a task-based loss, we improve over the state of the art by up to 2.67% labeled F1.

1 Introduction
Combinatory Categorial Grammar (CCG; Steedman, 2000) parsing is challenging due to its so-called “spurious” ambiguity that permits a large number of non-standard derivations (Vijay-Shanker and Weir, 1993; Kuhlmann and Satta, 2014). To address this, the de facto models resort to chart-based CKY (Hockenmaier, 2003; Clark and Curran, 2007), despite CCG being naturally compatible with shift-reduce parsing (Ades and Steedman, 1982). More recently, Zhang and Clark (2011) introduced the first shift-reduce model for CCG, which also showed substantial improvements over the long-established state of the art (Clark and Curran, 2007).

The success of the shift-reduce model (Zhang and Clark, 2011) can be tied to two main contributing factors. First, without any feature locality restrictions, it is able to use a much richer feature set; while intensive feature engineering is inevitable, it has nevertheless delivered an effective and conceptually simpler alternative for both parameter estimation and inference. Second, it couples beam search with global optimization (Collins, 2002; Collins and Roark, 2004; Zhang and Clark, 2008), which makes it less prone to search errors than fully greedy models (Huang et al., 2012).

In this paper, we present a neural architecture for shift-reduce CCG parsing based on long short-term memories (LSTMs; Hochreiter and Schmidhuber, 1997). Our model is inspired by Dyer et al. (2015), in which we explicitly linearize the complete history of parser states in an incremental fashion by requiring no feature engineering (Zhang and Clark, 2011; Xu et al., 2014), and no atomic feature sets (Chen and Manning, 2014). However, a key difference is that we achieve this linearization without relying on any additional control operations or compositional tree structures (Socher et al., 2010; Socher et al., 2011; Socher et al., 2013), both of which are vital in the architecture of Dyer et al. (2015). Crucially, unlike the sequence-to-sequence transduction model of Vinyals et al. (2015), which primarily conditions on the input words, our model is sensitive to all aspects of the parsing history, including arbitrary positions in the input.

As another contribution, we present a global LSTM parsing model by adapting an expected F-measure loss (Xu et al., 2016). As well as naturally incorporating beam search during training, this loss optimizes the model towards the final evaluation metric (Goodman, 1996; Smith and Eisner, 2006; Auli and Lopez, 2011b), allowing it to learn shift-reduce action sequences that lead to parses with high expected F-scores. We further show the globally optimized model can be leveraged with greedy inference, resulting in a deterministic parser as accurate
Our parser models normal-form derivations (Eisner, 1996) in CCGBank. However, unlike Zhang and Clark (2011), derivations are not restricted to be normal-form during inference.
longer time delays and makes them less susceptible to the vanishing gradient problem (Bengio et al., 1994). Hence, they are better at modeling temporal events that are arbitrarily far in a sequence.

Several extensions to the vanilla LSTM have been proposed over time, each with a modified instantiation of $\Phi_\theta$ that exerts refined control over e.g., whether the cell state could be reset (Gers et al., 2000) or whether extra connections are added to the cell state (Gers and Schmidhuber, 2000). Our instantiation is as follows for all LSTMs:

$$
\begin{align*}
i_t &= \sigma(W_{ix} x_t + W_{ih} h_{t-1} + W_{i,c} c_{t-1} + b_i) \\
f_t &= \sigma(W_{fx} x_t + W_{fh} h_{t-1} + W_{f,c} c_{t-1} + b_f) \\
c_t &= f_t \odot c_{t-1} + i_t \odot \tanh(W_{cx} x_t + W_{ch} h_{t-1} + b_c) \\
o_t &= \sigma(W_{ox} x_t + W_{oh} h_{t-1} + W_{o,c} c_t + b_o) \\
h_t &= o_t \odot \tanh(c_t),
\end{align*}
$$

where $\sigma$ is the sigmoid activation and $\odot$ is the element-wise product.

In addition to unidirectional LSTMs that model an input sequence $x_0, x_1, \ldots, x_{n-1}$ in a strict left-to-right order, we also use bidirectional LSTMs (Graves and Schmidhuber, 2005) (BLSTMs), which read the input from both directions with two independent LSTMs. At each step, the forward hidden state $h_t$ is computed using Eq. 2 for $t = (0, 1, \ldots, n - 1)$; and the backward hidden state $h_t$ is computed similarly but from the reverse direction for $t = (n - 1, n - 2, \ldots, 0)$. Together, the two hidden states at each step $t$ capture both past and future contexts, and the representation for each $x_t$ is obtained as the concatenation $[h_t; h_t]$.

### 3.2 Embeddings

The neural network model employed by Chen and Manning (2014), and followed by a number of other parsers (Weiss et al., 2015; Zhou et al., 2015; Ambati et al., 2016; Andor et al., 2016; Xu et al., 2016) allows higher-order feature conjunctions to be automatically discovered from a set of dense feature embeddings. However, a set of atomic feature templates, which are only sensitive to contexts from the top few elements on the stack and queue are still needed to dictate the choice of these embeddings. Instead, we dispense with such templates and seek to design a model that is sensitive to both local and non-local contexts, on both the stack and queue.

Consequently, embeddings represent atomic input units that are added to our parser and are preserved throughout parsing. In total we use four types of embeddings, namely, word, CCG category, POS and action, where each has an associated look-up table that maps a string of that type to its embedding. The look-up table for words is $\mathcal{L}_w \subseteq \mathbb{R}^{k \times |w|}$, where $k$ is the embedding dimension and $|w|$ is the size of the vocabulary. Similarly, we have look-up tables for CCG categories, $\mathcal{L}_c \subseteq \mathbb{R}^{l \times |c|}$, for the three types of actions, $\mathcal{L}_a \subseteq \mathbb{R}^{m \times 3}$, and for POS tags, $\mathcal{L}_p \subseteq \mathbb{R}^{n \times |p|}$.

### 3.3 Model

**Parser.** Fig. 2 shows the deduction system of our parser.$^2$ We denote each parse item as $(j, \delta, \beta, \Delta)$, where $j$ is the positional index of the word at the front of the queue, $\delta$ is the stack (with its top element $s_0$ to the right), and $\beta$ is the queue (with its top element $w_j$ to the left) and $\Delta$ is the set of CCG dependencies realized for the input consumed so far. Each item is also associated with a step indicator $t$, signifying the number of actions applied to it and the goal is reached in $2n - 1 + \mu$ steps, where $\mu$ is the total number of un actions. We also define each action in our parser as a 4-tuple $(\tau_t, c_t, w_{ct}, p_{w_{ct}})$, where $\tau_t \in \{\text{sh}, \text{re}, \text{un}\}$ for $t \geq 1$, $c_t$ is the resulting category of $\tau_t$, and $w_{ct}$ is the head word attached to

---

$^2$We assume greedy inference unless otherwise stated.
obtained as $w(\cdot)$ as the input (padded with a special embedding $g$) to $W$ with the complete parsing each sentence, we feed $W$ with the complete all aspects of the parsing history. Concretely, we facilitate this fact by completely linearizing our model as five LSTMs, comprising four unidirectional LSTMs and the bidirectional LSTM (right). The shaded cells on the left represent the hidden layer $V$, $X$ and $Y$ respectively.

LSTM model. LSTMs are designed to handle time-series data, in a purely sequential fashion, and we try to exploit this fact by completely linearizing all aspects of the parsing history. Concretely, we factor the model as five LSTMs, comprising four unidirectional ones, denoted as U, V, X and Y, and an additional BLSTM, denoted as W (Fig. 3). Before parsing each sentence, we feed W with the complete input (padded with a special embedding $g$ as the end of sentence token); and we use $w_j = [h_j^W; h_j^Y]$ to represent $w_j$ in subsequent steps.\(^4\) We also add $\perp$ to the other 4 unidirectional LSTMs as initialization.

Given this factorization, the stack representation for a parse item $(j, \delta, \beta, \Delta)$ at step $t$, for $t \geq 1$, is obtained as

\[
\delta_t = [h_t^U; h_t^V; h_t^X; h_t^Y],
\]

and together with $w_j$, $[\delta_t; w_j]$ gives a representation for the parse item. For the axiom item, we represent it as $[\delta_0; w_0]$, where $\delta_0 = [h_0^U; h_0^V; h_0^X; h_0^Y]$.

Each time the parser applies an action $(\tau_t, c_t, w_{ct}, p_{w_{ct}})$, we update the model by adding the embedding of $\tau_t$, denoted as $\mathcal{L}_a(\tau_t)$ onto U, and adding the other three embeddings of the action 4-tuple, that is, $\mathcal{L}_c(c_t)$, $\mathcal{L}_w(w_{ct})$ and $\mathcal{L}_p(p_{w_{ct}})$, onto V, X and Y respectively.

To predict the next action, we first derive an action hidden layer $b_t$, by passing the parse item representation $[\delta_t; w_j]$ through an affine transformation, s.t.

\[
b_t = f(B[\delta_t; w_j] + r),
\]

where $B$ is a parameter matrix of the model, $r$ is a bias vector and $f$ is a ReLU non-linearity (Nair and Hinton, 2010). Then we apply another affine transformation (with $A$ as the weights and $s$ as the bias) to $b_t$:

\[
a_t = f(AB_t + s),
\]

and obtain the probability of the $i$th action in $a_t$ as

\[
p(\tau_t^i|b_t) = \frac{\exp\{a_t^i\}}{\sum_{k \in \mathcal{T}(\delta_t, \beta_t)} \exp\{a_t^k\}},
\]

where $\mathcal{T}(\delta_t, \beta_t)$ is the set of feasible actions for the current parse item, and $\tau_t^i \in \mathcal{T}(\delta_t, \beta_t)$.

3.4 Derivations and Dependency Structures

Our model naturally linearizes CCG derivations “incrementally” following their post-order traversals. As such, the four unidirectional LSTMs always have the same number of steps; and at each step, the concatenation of their hidden states (Eq. 3) represents a point in a CCG derivation (i.e., an action 4-tuple).

Due to the large amount of flexibility in how dependencies are realized in CCG (Hockenmaier, 2003; Clark and Curran, 2007), and in line with most existing CCG parsing models, including dependency models, we have chosen to model CCG derivations, rather than dependency structures.\(^5\) We also hypothesize that tree structures are not necessary for the current model, since they are already implicit in the linearized derivations; similarly, we have found the action embeddings to be nonessential ($\S$5.2).

3.5 Training

As a baseline, we first train a greedy model, in which we maximize the log-likelihood of each target action in the training data. More specifically, let $(\tau_1, \ldots, \tau_T)$ be the gold-standard action sequence for a training sentence $n$, a cross-entropy criterion is used to obtain the error gradients, and for each sentence, training involves minimizing

\[
L(\theta) = -\log \prod_{t=1}^{T_n} p(\tau_t^g|b_t) = -\sum_{t=1}^{T_n} \log p(\tau_t^g|b_t),
\]

where $\theta$ is the set of all parameters in the model.

\(^3\)In case of multiple heads, we always choose the first one.

\(^4\)Word and POS embeddings are concatenated at each input position $j$, for $0 \leq j < n$; and $w_n = [h_n^Y; h_n^W]$.

\(^5\)Most CCG dependency models (e.g., see Clark and Curran (2007) and Xu et al. (2014)) model CCG derivations with dependency features.
As other greedy models (e.g., see Chen and Manning (2014) and Dyer et al. (2015)), our greedy model is locally optimized, and suffers from the label bias problem (Andor et al., 2016). A partial solution to this is to use beam search at test time, thereby recovering higher scoring action sequences that would otherwise be unreachable with fully greedy inference. In practice, this has limited effect (Table 2), and a number of more principled solutions have been recently proposed to derive globally optimized models during training (Watanabe and Sumita, 2015; Weiss et al., 2015; Zhou et al., 2015; Andor et al., 2016). Here, we extend our greedy model into a global one by adapting the expected F-measure loss of Xu et al. (2016). To our best knowledge, this is the first attempt to train a globally optimized LSTM shift-reduce parser.

Let $\theta = \{U, V, X, Y, W, B, A\}$ be the weights of the baseline greedy model, we initialize the weights of the global model, which has the same architecture as the baseline, to $\theta$, and we reoptimize $\theta$ in multiple training epochs as follows:

1. Pick a sentence $x_n$ from the training set, decode it with beam search, and generate a $k$-best list of output parses with the current $\theta$, denoted as $\Lambda(x_n)$.

2. For each parse $y_i$ in $\Lambda(x_n)$, compute its sentence-level F1 using the set of dependencies in the $\Delta$ field of its parse item. In addition, let $|y_i|$ be the total number of actions that derived $y_i$ and $s_\theta(y_i^j)$ be the softmax action score of the $j$th action, given by the LSTM model. Compute the log-linear score of its action sequence as $\rho(y_i) = \sum_{j=1}^{|y_i|} \log s_\theta(y_i^j)$.

3. Compute the negative expected F1 objective (defined below) for $x_n$ and minimize it using stochastic gradient descent (maximizing expected F1). Repeat these three steps for the remaining sentences.

More formally, the loss $J(\theta)$, is defined as

$$J(\theta) = -XF1(\theta) = - \sum_{y_i \in \Lambda(x_n)} \rho(y_i) \cdot F1(\Delta_{y_i}, \Delta^G_{x_n}),$$

where $F1(\Delta_{y_i}, \Delta^G_{x_n})$ is the sentence level F1 of the parse derived by $y_i$, with respect to the gold-standard dependency structure $\Delta^G_{x_n}$ of $x_n$; $\rho(y_i|\theta)$ is the normalized probability score of the action sequence $y_i$, computed as

$$\rho(y_i|\theta) = \frac{\exp\{\gamma \rho(y_i)\}}{\sum_{y \in \Lambda(x_n)} \exp\{\gamma \rho(y)\}}.$$

where $\gamma$ is a parameter that sharpens or flattens the distribution (Tromble et al., 2008). Different from the maximum-likelihood objective, XF1 optimizes the model on a sequence level and towards the final evaluation metric, by taking into account all action sequences in $\Lambda(x_n)$.

4 Attention-Based LSTM Supertagging

In addition to the size of the label space, supertagging is difficult because CCG categories can encode long-range dependencies and tagging decisions frequently depend on non-local contexts. For example, in He went to the zoo with a cat, a possible category for with, $(S'\backslash NP) \backslash (S'\backslash NP)/NP$, depends on the word went further back in the sentence.

Recently a number of RNN models have been proposed for CCG supertagging (Xu et al., 2015; Lewis et al., 2016; Vaswani et al., 2016; Xu et al., 2016), and such models show dramatic improvements over non-recurrent models (Lewis and Steedman, 2014b). Although the underlying models differ in their exact architectures, all of them make each tagging decision using only the hidden states at the current input position, and this imposes a potential bottleneck in the model. To mitigate this, we generalize the attention mechanisms of Bahdanau et al. (2015) and Luong et al. (2015), and adapt them to supertagging, by allowing the model to explicitly use hidden states from more than one input positions for tagging each word. Similar to Bahdanau et al. (2015) and Luong et al. (2015), a key feature in
our model is a soft alignment vector that weights the relative importance of the considered hidden states.

For an input sentence \( w_0, w_1, \ldots, w_{n-1} \), we consider \( w_t = [h_t; \tilde{h}_t] \) (3.1) to be the representation of the \( t \)th word (\( 0 \leq t < n \), \( w_t \in \mathbb{R}^{d_x} \)), given by a BLSTM with a hidden state size \( d \) for both its forward and backward layers.\(^9\) Let \( k \) be a context window size hyperparameter, we define \( H_t \in \mathbb{R}^{d \times (k-1)} \) as

\[
H_t = \left[ w_{t-k/2}, \ldots, w_{t-1}, w_{t+1}, \ldots, w_{t+k/2} \right],
\]

which contains representations for all words in the size \( k \) window except \( w_t \). At each position \( t \), the attention model derives a context vector \( c_t \in \mathbb{R}^{d \times 1} \) (defined below) from \( H_t \), which is used in conjunction with \( w_t \) to produce an attentional hidden layer:

\[
x_t = f(M[c_t; w_t] + m),
\]

where \( f \) is a ReLU non-linearity, \( M \in \mathbb{R}^{g \times kd} \) is a learned weight matrix, \( m \) is a bias term, and \( g \) is the size of \( x_t \). Then \( x_t \) is used to produce another hidden layer (with \( N \) as the weights and \( n \) as the bias):

\[
z_t = Nx_t + n,
\]

and the predictive distribution over categories is obtained by feeding \( z_t \) through a softmax activation.

In order to derive the context vector \( c_t \), we first compute \( b_t \in \mathbb{R}^{(k-1) \times 1} \) from \( H_t \) and \( w_t \) using \( \alpha \in \mathbb{R}^{1 \times d} \), s.t. the \( i \)th entry in \( b_t \) is

\[
b_t^i = \alpha[w_{T^i}; w_t],
\]

for \( i \in [0, k-1], T = \{ t-k/2, \ldots, t-1, t+1, \ldots, t+k/2 \} \); and \( c_t \) is derived as follows:

\[
a_t = \text{softmax}(b_t),
\]

\[
c_t = H_t a_t,
\]

where \( a_t \) is the alignment vector. We also experiment with two types of attention reminiscent of the global and local models in Luong et al. (2015), where the first attends over all input words (\( k = n \)) and the second over a local window.

It is worth noting that two other works have concurrently tackled supertagging with BLSTM models. In Vaswani et al. (2016), a language model layer is added on top of a BLSTM, which allows embeddings of previously predicted tags to propagate through and influence the pending tagging decision. However, the language model layer is only effective when both scheduled sampling for training (Bengio et al., 2015) and beam search for inference are used. We show our attention-based models match their performance, with only standard training and greedy decoding. Additionally, Lewis et al. (2016) presented a BLSTM model with two layers of stacking in each direction; and as an internal baseline, we show a non-stacking BLSTM without attention can achieve the same accuracy.

5 Experiments

Dataset and baselines. We conducted all experiments on CCGBank (Hockenmaier and Steedman, 2007) with the standard splits.\(^10\) We assigned POS tags with the C&C POS tagger, and used 10-fold jackknifing for both POS tagging and supertagging. All parsers were evaluated using F1 over labeled CCG dependencies.

For supertagging, the baseline models are the RNN model of Xu et al. (2015), the bidirectional RNN (BRNN) model of Xu et al. (2016), and the BLSTM supertagging models in Vaswani et al. (2016) and Lewis et al. (2016). For parsing experiments, we compared with the global beam-search shift-reduce parsers of Zhang and Clark (2011) and Xu et al. (2014). One neural shift-reduce CCG parser baseline is Ambati et al. (2016), which is a beam-search shift-reduce parser based on Chen and Manning (2014) and Weiss et al. (2015); and the others are the RNN shift-reduce models in Xu et al. (2016). Additionally, the chart-based C&C parser was included by default.

Model and training parameters.\(^11\) All our LSTM models are non-stacking with a single layer.\(^12\) For the supertagging models, the LSTM

\(^9\)Unlike in the parsing model, POS tags are excluded.

\(^10\)Training: Sections 02-21 (39,604 sentences). Development: Section 00 (1,913 sentences). Test: Section 23 (2,407 sentences).

\(^11\)We implemented all models using the CNN toolkit: https://github.com/clab/cnn.

\(^12\)The BLSTMs have a single layer in each direction. We experimented with 2 layers in all models during development and found negligible improvements.
hidden state size is 256, and the size of the attentional hidden layer ($x_t$, Eq. 5) is 200. All parsing model LSTMs have a hidden state size of 128, and the size of the action hidden layer ($b_t$, Eq. 4) is 80.

Pretrained word embeddings for all models are 100-dimensional (Turian et al., 2010), and all other embeddings are 50-dimensional. We also pretrained CCG lexical category and POS embeddings on the concatenation of the training data and a Wikipedia dump parsed with C&C. All other parameters were uniformly initialized in $\pm \sqrt{6/(r + c)}$, where $r$ and $c$ are the number of rows and columns of a matrix (Glorot and Bengio, 2010).

For training, we used plain non-minibatched stochastic gradient descent with an initial learning rate $\eta_0 = 0.1$ and we kept iterating in epochs until accuracy no longer increases on the dev set. For all models, a learning rate schedule $\eta_e = \eta_0/(1 + \lambda e)$, where $\lambda = 0.08$ was used for $e \geq 11$. Gradients were clipped whenever their norm exceeds 5. Dropout training as suggested by Zaremba et al. (2014), with a dropout rate of 0.3, and an $\ell_2$ penalty of $1 \times 10^{-5}$, were applied to all models.

### 5.1 Supertagging Results

Table 1 summarizes 1-best supertagging results. Our baseline BLSTM model without attention achieves the same level of accuracy as Lewis et al. (2016) and the baseline BLSTM model of Vaswani et al. (2016). Compared with the latter, our hidden state size is 50% smaller (256 vs. 512).

For training and testing the local attention model (BLSTM-local), we used an attention window size of 5 (tuned on the dev set), and it gives an improvement of 0.94% over the BRNN supertagger (Xu et al., 2016), achieving an accuracy on par with the beam-search (size 12) model of Vaswani et al. (2016) that is enhanced with a language model. Despite being able to consider wider contexts than the local model, the global attention model (BLSTM-global) did not show further gains, hence we used BLSTM-local for all parsing experiments below.

### 5.2 Parsing Results

All parsers we consider use a supertagger probability cutoff $\beta$ to prune categories less likely than $\beta$ times the probability of the best category in a distribution: for the C&C parser, it uses an adaptive strategy to backoff to smaller $\beta$ values if no spanning analysis is found given an initial $\beta$ setting; for all the shift-reduce parsers, fixed $\beta$ values are used without backing off. Since $\beta$ determines the derivation space of a parser, it has a large impact on the final parsing accuracy.

For the maximum-likelihood greedy model, we found using a small $\beta$ value (bigger ambiguity) for training significantly improved accuracy, and we chose $\beta = 1 \times 10^{-5}$ (5.22 categories per word with jackknifing) via development experiments. This reinforces the findings in a number of other CCG parsers (Clark and Curran, 2007; Auli and Lopez, 2011a): even though a smaller $\beta$ increases ambiguity, it leads to more accurate models at test time. On the other hand, we found using larger $\beta$ values at test time led to significantly better results (Table 2). And this differs from the beam-search models that use the same $\beta$ value for both training and testing (Zhang and Clark, 2011; Xu et al., 2014).

---

<table>
<thead>
<tr>
<th>Model</th>
<th>Dev</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>C&amp;C</td>
<td>91.50</td>
<td>92.02</td>
</tr>
<tr>
<td>Xu et al. (2015)</td>
<td>93.07</td>
<td>93.00</td>
</tr>
<tr>
<td>Xu et al. (2016)</td>
<td>94.10</td>
<td>94.35</td>
</tr>
<tr>
<td>Lewis et al. (2016)</td>
<td>94.11</td>
<td>94.32</td>
</tr>
<tr>
<td>Vaswani et al. (2016)</td>
<td>94.08</td>
<td>-</td>
</tr>
<tr>
<td>Vaswani et al. (2016) +LM +beam</td>
<td>94.24</td>
<td>94.50</td>
</tr>
<tr>
<td>BLSTM</td>
<td>94.11</td>
<td>94.29</td>
</tr>
<tr>
<td>BLSTM-local</td>
<td><strong>94.31</strong></td>
<td><strong>94.46</strong></td>
</tr>
<tr>
<td>BLSTM-global</td>
<td>94.22</td>
<td>94.42</td>
</tr>
</tbody>
</table>

**Table 1:** 1-best supertagging results on both the dev and test sets. BLSTM is the baseline model without attention; BLSTM-local and -global are the two attention-based models.

---

<table>
<thead>
<tr>
<th>Supertagger $\beta$</th>
<th>Beam</th>
<th>0.09</th>
<th>0.07</th>
<th>0.06</th>
<th>0.01</th>
<th>0.001</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>86.49</td>
<td>86.52</td>
<td>86.56</td>
<td>86.26</td>
<td>85.80</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>86.55</td>
<td>86.58</td>
<td>86.63</td>
<td>86.39</td>
<td>86.01</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>86.61</td>
<td>86.64</td>
<td><strong>86.67</strong></td>
<td>86.40</td>
<td>86.07</td>
<td></td>
</tr>
</tbody>
</table>

**Table 2:** Tuning beam size and supertagger $\beta$ on the dev set.

<table>
<thead>
<tr>
<th>Model</th>
<th>LP</th>
<th>LR</th>
<th>LF</th>
<th>CAT</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSTM-w</td>
<td><strong>90.13</strong></td>
<td>76.99</td>
<td>83.05</td>
<td>94.24</td>
</tr>
<tr>
<td>LSTM-w+c</td>
<td>89.37</td>
<td>83.25</td>
<td>86.20</td>
<td>94.34</td>
</tr>
<tr>
<td>LSTM-w+c+a</td>
<td>89.31</td>
<td>83.39</td>
<td>86.25</td>
<td>94.38</td>
</tr>
<tr>
<td>LSTM-w+c+a+p</td>
<td>89.43</td>
<td><strong>83.86</strong></td>
<td><strong>86.56</strong></td>
<td><strong>94.47</strong></td>
</tr>
</tbody>
</table>

**Table 3:** F1 on dev for all the greedy models.

---

13 We used the gensim word2vec toolkit: https://radimrehurek.com/gensim/.
The greedy model. Table 3 shows the dev set results for all greedy models, where the four types of embeddings, that is, word (w), CCG category (c), action (a) and POS (p), are gradually introduced. The full model LSTM-w+c+a+p surpasses all previous shift-reduce models (Table 4), achieving a dev set accuracy of 86.56%. Category embeddings (LSTM-w+c) yielded a large gain over using word embeddings alone (LSTM-w); action embeddings (LSTM-w+c+a) provided little improvement, but further adding POS embeddings (LSTM-w+c+a+p) gave noticeable recall (+0.61%) and F1 improvements (+0.36%) over LSTM-w+c. Fig. 4a shows the learning curves, where all models converged in under 30 epochs.

The XF1 model. Table 4 also shows the results for the XF1 models (LSTM-XF1), which use all four types of embeddings. We used a beam size of 8, and a $\beta$ value of 0.06 for both training and testing (tuned on the dev set); and training took 12 epochs to converge (Fig. 4b), with an F1 of 87.45% on the dev set. Decoding the XF1 model with greedy inference only slightly decreased recall and F1, and this resulted in a highly accurate deterministic parser. On the test set, our XF1 greedy model gives 2.67% F1 improvement over the greedy model in Xu et al. (2016); and the beam-search XF1 model achieves an F1 improvement of 1.34% compared with the XF1 model of Xu et al. (2016).

Table 4: Parsing results on the dev (Section 00) and test (Section 23) sets with 100% coverage, with all LSTM models using the BLSTM-local supertagging model. All experiments using auto POS. CAT (lexical category assignment accuracy). LSTM-greedy is the full greedy parser.

Table 5: Effect of different supertaggers on the full greedy parser. LSTM-greedy is the same parser as in Table 4, which uses the BLSTM-local supertagger.

<table>
<thead>
<tr>
<th>Model</th>
<th>Dev</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSTM-FLASH</td>
<td>85.86</td>
<td>86.37</td>
</tr>
<tr>
<td>LSTM-BLSTM</td>
<td>86.06</td>
<td>86.64</td>
</tr>
<tr>
<td>LSTM-greedy</td>
<td>86.56</td>
<td>86.83</td>
</tr>
</tbody>
</table>

Figure 4: Learning curves with dev F-scores for all models.
Model | LP  | LR  | LF  
---|-----|-----|-----
Xu et al. (2015) | 87.68 | 86.41 | 87.04 
Lewis et al. (2016) | 87.7 | 86.7 | 87.2 
Lewis et al. (2016)* | 88.6 | 87.5 | 88.1 
Vaswani et al. (2016)* | - | - | 88.32 
Lee et al. (2016) | - | - | 88.7 
LSTM-XF1 (beam = 1) | 89.85 | 85.51 | 87.62 
LSTM-XF1 (beam = 8) | 89.81 | 85.81 | 87.76 

Table 6: Comparison of our XF1 models with chart-based parsers on the test set. * denotes a tri-trained model and † indicates a different POS tagger.

Effect of the supertagger. To isolate the parsing model from the supertagging model, we first experimented with the BRNN supertagging model as in Xu et al. (2016) for both training and testing the full greedy LSTM parser. Using this supertagger, we still achieved the highest F1 (85.86%) on the dev set (LSTM-BRNN, Table 5) in comparison with all previous shift-reduce models; and an improvement of 1.42% F1 over the greedy model of Xu et al. (2016) was obtained on the test set (Table 4). We then experimented with using the baseline BLSTM supertagging model for parsing (LSTM-BLSTM), and observed the attention-based setup (LSTM-greedy) outperformed it, despite the attention-based supertagger (BLSTM-local) did not give better multi-tagging accuracy. We owe this to the fact that large β cutoff values—resulting in almost deterministic supertagging decisions on average—are required by the parser during inference; for instance, BLSTM-local has an average ambiguity of 1.09 on the dev set when β = 0.06.14

Comparison with chart-based models. For completeness and to put our results in perspective, we compare our XF1 models with other CCG parsers in the literature (Table 6): Xu et al. (2015) is the log-linear C&C dependency hybrid model with an RNN supertagger front-end; Lewis et al. (2016) is an LSTM supertagger-factored parser using the A* CCG parsing algorithm of Lewis and Steedman (2014a); Vaswani et al. (2016) combine a BLSTM supertagger with a new version of the C&C parser (Clark et al., 2015) that uses a max-violation perceptron, which significantly improves over the original C&C models; and finally, a global recursive neural network model with A* decoding (Lee et al., 2016). We note that all these alternative models—with the exception of Xu et al. (2015) and Lewis et al. (2016)—use structured training that accounts for violations of the gold-standard, and we conjecture further improvements for our model are possible by incorporating such mechanisms.15

6 Conclusion

We have presented an LSTM parsing model for CCG, with a factorization allowing the linearization of the complete parsing history. We have shown that this simple model is highly effective, with results outperforming all previous shift-reduce CCG parsers. We have also shown global optimization benefits an LSTM shift-reduce model; and contrary to previous findings with the averaged perceptron (Zhang and Clark, 2008), we empirically demonstrated beam-search inference is not necessary for our globally optimized model. For future work, a natural direction is to explore integrated supertagging and parsing in a single neural model (Zhang and Weiss, 2016).

Acknowledgment

I acknowledge the support from the Carnegie Trust for the Universities of Scotland through a Carnegie Scholarship.

References

Michael Auli and Adam Lopez. 2011a. A comparison of loopy belief propagation and dual decomposition for 14All β cutoffs were tuned on the dev set; for BRNN, we found the same β settings as in Xu et al. (2016) to be optimal; for BLSTM, β = 4 × 10−5 for training (with an ambiguity of 5.27) and β = 0.02 for testing (with an ambiguity of 1.17).

15Our XF1 training considers shift-reduce action sequences, but not violations of the gold-standard (e.g., see Huang et al. (2012), Watanabe and Sumita (2015), Zhou et al. (2015) and Andor et al. (2016)).
integrated CCG supertagging and parsing. In *Proc. of ACL*.


Felix Gers and Jürgen Schmidhuber. 2000. Recurrent nets that time and count. In *Neural Networks*. IEEE.


Richard Socher, Christopher Manning, and Andrew Ng. 2010. Learning continuous phrase representations and syntactic parsing with recursive neural networks. In


