What's Going On in Neural Constituency Parsers? An Analysis

David Gaddy, Mitchell Stern, and Dan Klein
Computer Science Division
University of California, Berkeley
{dgaddy,mitchell,klein}@berkeley.edu

Abstract

A number of differences have emerged between modern and classic approaches to constituency parsing in recent years, with structural components like grammars and feature-rich lexicons becoming less central while recurrent neural network representations rise in popularity. The goal of this work is to analyze the extent to which information provided directly by the model structure in classical systems is still being captured by neural methods. To this end, we propose a high-performance neural model (92.08 F1 on PTB) that is representative of recent work and perform a series of investigative experiments. We find that our model implicitly learns to encode much of the same information that was explicitly provided by grammars and lexicons in the past, indicating that this scaffolding can largely be subsumed by powerful general-purpose neural machinery.

1 Introduction

In the past several years, many aspects of constituency parsing and natural language processing in general have changed. Grammars, which were once the central component of many parsers, have played a continually decreasing role. Rich lexicons and handcrafted lexical features have become less common as well. On the other hand, recurrent neural networks have gained traction as a powerful and general purpose tool for representation. So far, not much has been shown about how neural networks are able to compensate for the removal of the structures used in past models. To gain insight, we introduce a parser that is representative of recent trends and analyze its learned representations to determine what information it captures and what is important for its strong performance.

Our parser is a natural extension of recent work in constituency parsing. We combine a common span representation based on recurrent neural networks with a novel, simplified scoring model. In addition, we replace the externally predicted part-of-speech tags used in some recent systems with character-level word representations. Our parser achieves a test F1 score of 92.08 on section 23 of the Penn Treebank, exceeding the performance of many other state-of-the-art models evaluated under comparable conditions. Section 2 describes our model in detail.

The remainder of the paper is focused on analysis. In Section 3, we look at the decline of grammars and output correlations. Past work in constituency parsing used context-free grammars with production rules governing adjacent labels (or more generally production-factored scores) to propagate information and capture correlations between output decisions (Collins, 1997; Charniak and Johnson, 2005; Petrov and Klein, 2007; Hall et al., 2014). Many recent parsers no longer have explicit grammar production rules, but still use information about other predictions, allowing them to capture output correlations (Dyer et al., 2016; Choe and Charniak, 2016). Beyond this, there are some parsers that use no context for bracket scoring and only include mild output correlations in the form of tree constraints (Cross and Huang, 2016b; Stern et al., 2017). In our experiments, we find that we can accurately predict parents from the representation given to a child. Since a simple classifier can predict the information provided by parent-child relations, this explains why the information no longer needs to be specified explicitly. We also show that we can completely remove output correlations from our model with a variant of our parser that makes independent span label decisions without any tree constraints while maintaining high F1 scores and mostly producing trees.

In Section 4, we look at lexical representations. In the past, parsers used a variety of cus-

tom lexical representations, such as word shape features, prefixes, suffixes, and special tokens for categories like numerals (Klein and Manning, 2003; Petrov and Klein, 2007; Finkel et al., 2008). Character-level models have shown promise in parsing and other NLP tasks as a way to remove the complexity of these lexical features (Ballesteros et al., 2015; Ling et al., 2015b; Kim et al., 2016; Coavoux and Crabbé, 2017; Liu and Zhang, 2017). We compare the performance of character-level representations and externally predicted part-of-speech tags and show that these two sources of information seem to fill a similar role. We also perform experiments showing that the representations learned with character-level models contain information that was hand-specified in some other models.

Finally, in Section 5 we look at the surface context captured by recurrent neural networks. Many recent parsers use LSTMs, a popular type of recurrent neural network, to combine and summarize context for making decisions (Choe and Charniak, 2016; Cross and Huang, 2016a; Dyer et al., 2016; Stern et al., 2017). Before LSTMs became common in parsing, systems that included surface features used a fixed-size window around the fenceposts at each end of a span (Charniak and Johnson, 2005; Finkel et al., 2008; Hall et al., 2014; Durrett and Klein, 2015), and the inference procedure handled most of the propagation of information from the rest of the sentence. We perform experiments showing that LSTMs capture far-away surface context and that this information is important for our parser's performance. We also provide evidence that word order of the far-away context is important and that the amount of context alone does not account for all of the gains seen with LSTMs.

Overall, we find that the same sources of information that were effective for grammar-driven parsers are also captured by parsers based on recurrent neural networks.

2 Parsing Model

In this section, we propose a span-based parsing model that combines components from several recent neural architectures for constituency parsing and other natural language tasks. While this system is primarily introduced for the purpose of our analysis, it also performs well as a parser in its own right, exhibiting some gains over comparable work. Our model is in many respects similar to the chart parser of Stern et al. (2017), but features a number of simplifications and improvements.

2.1 Overview

Abstractly, our model consists of a single scoring function $s(i, j, \ell)$ that assigns a real-valued score to every label ℓ for each span (i, j) in an input sentence. We take the set of available labels to be the collection of all nonterminals and unary chains observed in the training data, treating the latter as atomic units. The score of a tree T is defined as a sum over internal nodes of labeled span scores:

$$s(T) = \sum_{(i,j,\ell) \in T} s(i, j, \ell).$$

We note that, in contrast with many other chart parsers, our model can directly score n-ary trees without the need for binarization or other tree transformations. Under this setup, the parsing problem is to find the tree with the highest score:

$$\hat{T} = \operatorname*{argmax}_{T} s(T).$$

Our concrete implementation of $s(i, j, \ell)$ can be broken down into three pieces: word representation, span representation, and label scoring. We discuss each of these in turn.

2.2 Word Representation

One popular way to represent words is the use of word embeddings. We have a separate embedding for each word type in the training vocabulary and map all unknown words at test time to a single `<UNK>` token. In addition to word embeddings, character-level representations have also been gaining traction in recent years, with common choices including recurrent, convolutional, or bag-of-n-gram representations. These alleviate the unknown word problem by working with smaller, more frequent units, and readily capture morphological information not directly accessible through word embeddings. Character LSTMs in particular have proved useful in constituency parsing (Coavoux and Crabbé, 2017), dependency parsing (Ballesteros et al., 2015), part-of-speech tagging (Ling et al., 2015a), named entity recognition (Lample et al., 2016), and machine translation (Ling et al., 2015b), making them a natural choice for our system. We obtain a character-level representation for a word by running it through a bidirectional character LSTM and concatenating the final forward and backward outputs.

The complete representation of a given word is the concatenation of its word embedding and its character LSTM representation. While past work has also used sparse indicator features (Finkel et al., 2008) or part-of-speech tags predicted by an external system (Cross and Huang, 2016b) for additional word-level information, we find these to be unnecessary in the presence of a robust character-level representation.

2.3 Span Representation

To build up to spans, we first run a bidirectional LSTM over the sequence of word representations for an input sentence to obtain context-sensitive forward and backward representations $\mathbf{f}_i$ and $\mathbf{b}_i$ for each fencepost i. We then follow past work in dependency parsing (Wang and Chang, 2016) and constituency parsing (Cross and Huang, 2016b; Stern et al., 2017) in representing the span (i, j) by the concatenation of the corresponding forward and backward span differences:

$$\mathbf{r}_{ij} = [\mathbf{f}_j - \mathbf{f}_i, \mathbf{b}_i - \mathbf{b}_j].$$

See Figure 1 for an illustration.

2.4 Label Scoring

Finally, we implement the label scoring function by feeding the span representation through a one-layer feedforward network whose output dimensionality equals the number of possible labels. The score of a specific label ℓ is the corresponding component of the output vector:

$$s(i, j, \ell) = [\mathbf{W}_2\, g(\mathbf{W}_1 \mathbf{r}_{ij} + \mathbf{z}_1) + \mathbf{z}_2]_\ell\,,$$

where g is an elementwise ReLU nonlinearity.

2.5 Inference

Even though our model operates on n-ary trees, we can still employ a CKY-style algorithm for efficient globally optimal inference by introducing an auxiliary empty label $\varnothing$ with $s(i, j, \varnothing) = 0$ for all (i, j) to handle spans that are not constituents. Under this scheme, every binarization of a tree with empty labels at intermediate dummy nodes will have the same score, so an arbitrary binarization can be selected at training time with no effect on learning. We contrast this with the chart parser of Stern et al. (2017), which assigns different scores to different binarizations of the same underlying tree and in theory may exhibit varying performance depending on the method chosen for conversion.

With this change in place, let $s_{\text{best}}(i, j)$ denote the score of the best subtree spanning (i, j). For spans of length one, we need only consider the choice of label:

$$s_{\text{best}}(i, i+1) = \max_\ell s(i, i+1, \ell).$$

For general spans (i, j), we have the following recursion:

$$s_{\text{best}}(i, j) = \max_\ell s(i, j, \ell) + \max_k \left[s_{\text{best}}(i, k) + s_{\text{best}}(k, j) \right].$$

That is, we can independently select the best label for the current span and the best split point, where the score of a split is the sum of the best scores for the corresponding subtrees.

To parse the full sentence, we compute $s_{\text{best}}(0, n)$ using a bottom-up chart decoder, then traverse backpointers to recover the tree achieving that score. Nodes assigned the empty label are omitted during the reconstruction process to obtain the full n-ary tree. The overall complexity of this approach is $\mathcal{O}(n^3 + Ln^2)$, where n is the number of words and L is the total number of labels. We note that because our system does not use a grammar, there is no constant for the number of grammar rules multiplying the $\mathcal{O}(n^3)$ term as in traditional CKY parsing. In practice, the $\mathcal{O}(n^2)$ evaluations of the span scoring function corresponding to the $\mathcal{O}(Ln^2)$ term dominate runtime.

2.6 Training

As is common for structured prediction problems (Taskar et al., 2005), we use margin-based training to learn a model that satisfies the constraints

$$s(T^*) \geq s(T) + \Delta(T, T^*)$$

for each training example, where T^* denotes the gold output, T ranges over all valid trees, and Δ is the Hamming loss on labeled spans. Our training objective is the hinge loss:

$$\max\left(0,\, \max_T \left[s(T) + \Delta(T, T^*) \right] - s(T^*)\right).$$

This is equal to 0 when all constraints are satisfied, or the magnitude of the largest margin violation otherwise.

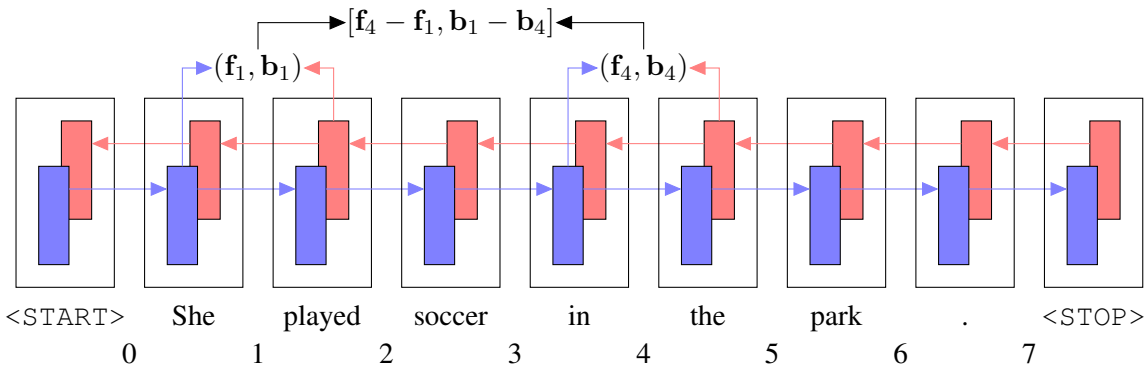

Figure 1: Span representations are computed by running a bidirectional LSTM over the input sentence and taking differences of the output vectors at the two endpoints. Here we illustrate the process for the span $(1, 4)$ corresponding to "played soccer in" in the example sentence.

Since Δ decomposes over spans, the inner loss-augmented decode $\max_T [s(T) + \Delta(T, T^*)]$ can be performed efficiently using a slight modification of the dynamic program used for inference. In particular, we replace $s(i, j, \ell)$ with $s(i, j, \ell) + 1[\ell \neq \ell^*_{ij}]$, where ℓ^*_{ij} is the label of span (i, j) in the gold tree T^*.

2.7 Results

We use the Penn Treebank (Marcus et al., 1993) for our experiments with the standard splits of sections 2-21 for training, section 22 for development, and section 23 for testing. Details about our model hyperparameters and training prodecure can be found in Appendix A.

Across 10 trials, our model achieves an average development F1 score of 92.22 on section 22 of the Penn Treebank. We use this as our primary point of comparison in all subsequent analysis. The model with the best score on the development set achieves a test F1 score of 92.08 on section 23 of the Penn Treebank, exceeding the performance of other recent state-of-the-art discriminative models which do not use external data or ensembling.[1]

3 Output Correlations

Output correlations are information about compatibility between outputs in a structured prediction model. Since outputs are all a function of the input, output correlations are not necessary for prediction when a model has access to the entire input. In practice, however, many models throughout NLP have found them useful (Collins, 1997; Lafferty et al., 2001; Koo and Collins, 2010), and Liang et al. (2008) provides theoretical results suggesting they may be useful for learning efficiently. In constituency parsing, there are two primary forms of output correlation typically captured by models. The first is correlations between label decisions, which often are captured by either production scores or the history in an incremental tree-creation procedure. The second, more subtle correlation comes from the enforcement of tree constraints, since the inclusion of one bracket can affect whether or not another bracket can be present. We explore these two classes of output correlations in Sections 3.1 and 3.2 below.

3.1 Parent Classification

The base parser introduced in Section 2 scores labeled brackets independently then uses a dynamic program to select a set of brackets that forms the highest-scoring tree. This independent labeling is an interesting departure from classical parsing work where correlations between predicted labels played a central role. It is natural to wonder why modeling label correlations isn't as important as it once was. Is there something about the neural representation that allows us to function without it? One possible explanation is that the neural machinery, in particular the LSTM, is handling much of the reconciliation between labels that was previously handled by an inference procedure. In other words, instead of using local information to suggest several brackets and letting the grammar handle interactions between them, the LSTM may be making decisions about brackets already in its latent state, allowing it to use the result of these decisions to inform other bracketings.

One way to explore this hypothesis would be

[1] Code for our parser is available at https://github.com/dgaddy/parser-analysis.

to evaluate whether the parser's learned representations could be used to predict parent labels of nodes in the tree. If the label of a node's parent can be predicted with high accuracy from the representation of its span, then little of the information about parent-child relations provided explicitly by a grammar has been lost. For this experiment, we freeze the input and LSTM parameters of our base model and train a new label scoring network to predict the label of a span's parent rather than the label of the span itself. We only predict parent labels for spans that have a bracket in the gold tree, so that all but the top level spans will have non-empty labels. The new network is trained with a margin loss.

After training on the standard training sections of the treebank, the network was able to correctly predict 92.3% of parent labels on the development set. This is fairly accurate, which supports the hypothesis that the representation knows a substantial amount about surrounding context in the output tree. For comparison, given only a span's label, the best you can do for predicting the parent is 43.3% with the majority class conditioned on the current label.

3.2 Independent Span Decisions

Like other recent parsers that do not capture correlations between output labels (Cross and Huang, 2016b; Stern et al., 2017), our base parser still does have some output correlations captured by the enforcement of tree constraints. In this section, we set out to determine the importance of these output correlations by making a version of the parser where they are removed. Although parsers are typically designed to form trees, the bracketing F1 measure used to evaluate parsers is still defined on non-tree outputs. To remove all output correlations from our parser, we can simply remove the tree constraint and independently make decisions about whether to include a bracketed span. The architecture is identical to the one described in Section 2, producing a vector of label scores for each span. We choose the label with the maximum score as the label for a span. As before, we fix the score of the empty label at zero, so if all other label scores are negative, the span will be left out of the set of predicted brackets. We train with independent margin losses for each span.

Ignoring tree well-formedness, the development F1 score of this independent span selection parser is 92.20, effectively matching the performance of the tree-constrained parser. In addition, we find that 94.5% of predicted bracketings for development set examples form valid trees, even though we did not explicitly encourage this. This high performance shows that our parser can function well even without modeling any output correlations.

4 Lexical Representation

In this section, we investigate several common choices for lexical representations of words and their role in neural parsing.

4.1 Alternate Word Representations

We compare the performance of our base model, which uses word embeddings and a character LSTM, with otherwise identical parsers that use other combinations of lexical representations. The results of these experiments are summarized in Table 1. First, we remove the character-level representations from our model, leaving only the word embeddings. We find that development performance drops from 92.22 F1 to 91.44 F1, showing that word embeddings alone do not capture sufficient information for state-of-the-art performance. Then, we replace the character-level representations with embeddings of part-of-speech tags predicted by the Stanford tagger (Toutanova et al., 2003). This model achieves a comparable development F1 score of 92.09, but unlike our base model relies on outputs from an external system. Next, we train a model which includes all three lexical representations: word embeddings, character LSTM representations, and part-of-speech tag embeddings. We find that development performance is nearly identical to the base model at 92.24 F1, suggesting that character representations and predicted part-of-speech tags provide much of the same information. Finally, we remove word embeddings and rely completely on character-level embeddings. After retuning the character LSTM size, we find that a slightly larger character LSTM can make up for the loss in word-level embeddings, giving a development F1 of 92.24.

4.2 Predicting Word Features

Past work in constituency parsing has demonstrated that indicator features on word shapes, suffixes, and similar attributes provide useful infor-

Word and Character LSTM	92.22
Word Only	91.44
Word and Tag	92.09
Word, Tag, and Character LSTM	92.24
Character Only	92.24

Table 1: Development F1 scores on section 22 of the Penn Treebank for different lexical representations.

mation beyond the identity of a word itself, especially for rare and unknown tokens (Finkel et al., 2008; Hall et al., 2014). We hypothesize that the character-level LSTM in our model learns similar information without the need for manual supervision. To test this, we take the word representations induced by the character LSTM in our parser as fixed word encodings, and train a small feed-forward network to predict binary word features defined in the Berkeley Parser (Petrov and Klein, 2007). We randomly split the vocabulary of the Penn Treebank into two subsets, using 80% of the word types for training and 20% for testing.

We find that the character LSTM representations allow for previously handcrafted indicator features to be predicted with accuracies of 99.7% or higher in all cases. The fact that this simple classifier performs so well indicates that the information contained in these features is readily available from our model's character-level encodings. A detailed breakdown of accuracy by feature can be found in Appendix B.

5 Context in the Sentence LSTM

In this section, we analyze where the information in the sentence-level LSTM hidden vectors comes from. Since the LSTM representations we use to make parsing decisions come from the fenceposts on each side of a span, we would like to understand whether they only capture information from the immediate vicinity of the fenceposts or if they also contain more distant information. Although an LSTM is theoretically capable of incorporating an arbitrarily large amount of context, it is unclear how much context it actually captures and whether this context is important for parsing accuracy.

5.1 Derivative Analysis

First, we would like to know if the LSTM features capture distant information. For this experiment, we use derivatives as a measure of sensitivity to changes in an input. If the derivative of a value with respect to a particular input is high, then that input has a large impact on the final value. For a particular component of an LSTM output vector, we compute its gradient with respect to each LSTM input vector, calculate the ℓ_2-norms of the gradients, and bucket the results according to distance from the output position. This process is repeated for every output position of each sentence in the development set, and the results are averaged within each bucket. Due to the scale of the required computation, we only use a subset of the output vector components to compute the average, sampling one at random per output vector.

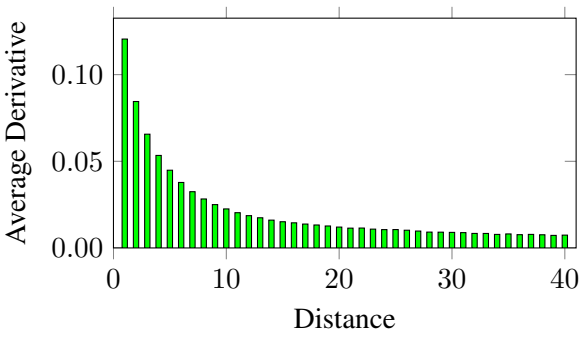

Figure 2: Average derivative of the LSTM output with respect to its input as a function of distance. The output is most sensitive to the closest words, but the tail of the distribution is fairly heavy, indicating that far-away words also have substantial impact.

Figure 2 illustrates how the average gradient norm is affected by the distance between the LSTM input and output. As would be expected, the closest input vectors have the largest effect on the hidden state. However, the tail of values is fairly heavy, with substantial gradient norms even for inputs 40 words away. This shows that far-away inputs do have an effect on the LSTM representation.

5.2 Truncation Analysis

Next, we investigate whether information in the LSTM representation about far-away inputs is actually important for parsing performance. To do so, we remove distant context information from our span encoding, representing spans by features obtained from LSTMs that are run on fixed-sized windows of size k around each fencepost. Figure 3 illustrates this truncated representation. Since the truncated representation also removes information about the size and position of the span in addition to the context words, we learn a position-dependent cell state initialization for each of the

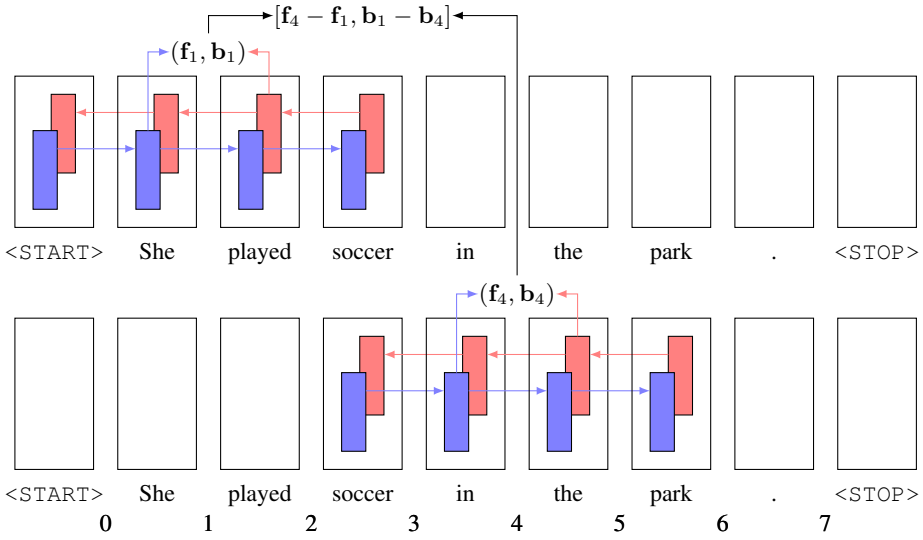

Figure 3: An example of creating a truncated span representation for the span "played soccer in" with context size $k = 2$. This representation is used to investigate the importance of information far away from the fenceposts of a span.

two LSTM directions to give a more fair comparison to the full LSTM. The use of a fixed-sized context window is reminiscent of prior work by Hall et al. (2014) and Durrett and Klein (2015), but here we use an LSTM instead of sparse features. We train parsers with different values of k and observe how their performance varies. All other architecture details and hyperparameters are the same as for the original model.

The blue points in Figure 4 show how the context size k affects parser performance for $k \in \{2, 3, 5, 10, 20, 30\}$. As with the derivative analysis, although most of the weight is carried by the nearby inputs, a nontrivial fraction of performance is due to context more than 10 words away.

5.3 Word Order

Now that we have established that long-distance information is important for parsing performance, we would like to know whether the order of the far-away words is important. Is the LSTM capturing far-away structure, or is the information more like a bag-of-words representation summarizing the words that appear?

To test the importance of order, we train a parser where information about the order of far-away words is destroyed. As illustrated in Figure 5, we run a separate LSTM over the entire sentence for each fencepost, shuffling the input depending on the particular fencepost being represented. We randomly shuffle words outside a context window

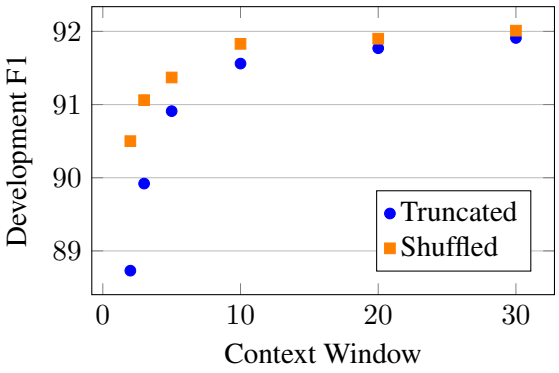

Figure 4: Development F1 as the amount of context given to the sentence-level LSTM varies. The blue points represent parser performance when the LSTM is truncated to a window around the fenceposts, showing that far-away context is important. The orange points represent performance when the full context is available but words outside a window around the fenceposts are shuffled, showing that the order of far-away context is also important.

of size k around the fencepost of interest, keeping words on the left and the right separate so that directional information is preserved but exact positions are lost.

The orange points in Figure 4 show the performance of this experiment with different context sizes k. We observe that including shuffled distant words is substantially better than truncating them completely. On the other hand, shuffling does cause performance to degrade relative to the base parser even when the unshuffled win-

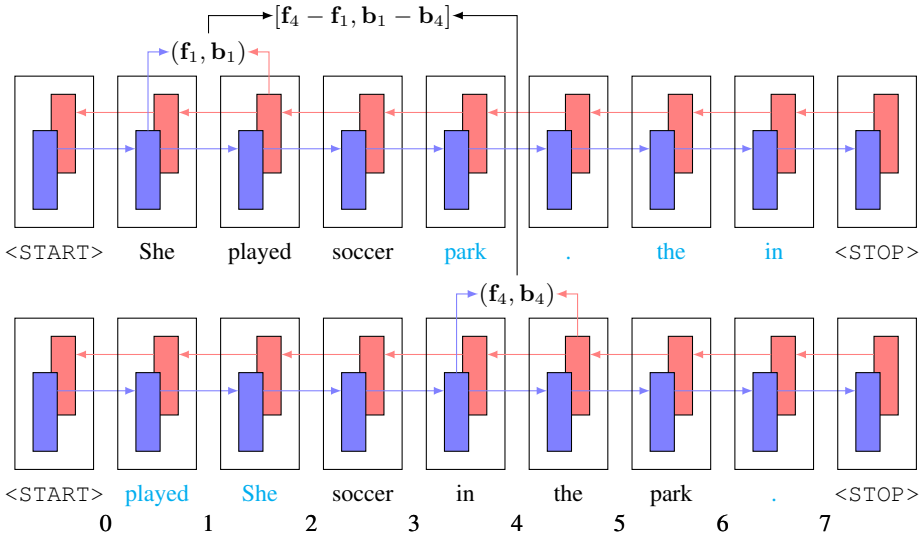

Figure 5: An example of creating a shuffled span representation for the span "played soccer in" with context size $k = 2$. The light blue words are outside the context window and are shuffled randomly. Shuffled representations are used to explore whether the order of far-away words is important.

dow is moderately large, indicating that the LSTM is propagating information that depends on the order of words in far-away positions.

5.4 LSTMs vs. Feedforward

Finally, we investigate whether the LSTM architecture itself is important for reasons other than just the amount of context it can capture. Like any architecture, the LSTM introduces particular inductive biases that affect what gets learned, and these could be important for parser performance. We run a version of the truncation experiment from Section 5.2 where we use a feedforward network in place of a sentence-level LSTM to process the surrounding context of each fencepost. The input to the network is the concatenation of the word representations that would be used as inputs for the truncated LSTM, and the output is a vector of the same size as the LSTM-based representation. As in Section 5.2, we wish to give our representation information about span size and position, so we also include a learned fencepost position embedding in the concatenated inputs to the network. We focus on context window size $k = 3$ for this experiment. We search among networks with one, two, or three hidden layers that are one, two, or four times the size of the LSTM hidden state.

Of all the feedforward networks tried, the maximum development performance was 83.39 F1, compared to 89.92 F1 for the LSTM-based truncation. This suggests that some property of the LSTM makes it better suited for the task of summarizing context than a flat feedforward network.

6 Related Analysis Work

Here we review other works that have performed similar analyses to ours in parsing and other areas of NLP. See Section 2 for a description of how our parser is related to other parsers.

Similar to our independent span prediction in Section 3.2, several works have found that their models still produce valid outputs for the majority of inputs even after relaxing well-formedness constraints. In dependency parsing, Zhang et al. (2017) and Chorowski et al. (2016) found that selecting dependency heads independently often resulted in valid trees for their parsers (95% and 99.5% of outputs form trees, respectively). In constituency parsing, the parser of Vinyals et al. (2015), which produced linearized parses token by token, was able to output valid constituency trees for the majority of sentences (98.5%) even though it was not constrained to do so.

Several other works have investigated what information is being captured within LSTM representations. Chawla et al. (2017) performed analysis of bidirectional LSTM representations in the context of named entity recognition. Although they were primarily interested in finding specific word types that were important for making decisions, they also analyzed how distance affected a word's impact. Shi et al. (2016) and Linzen et al.

(2016) perform analysis of LSTM representations in machine translation and language modeling respectively to determine whether syntactic information is present. Some of their techniques involve classification of features from LSTM hidden states, similar to our analysis in Sections 3.1 and 4.2.

In Section 5.4, we found that replacing an LSTM with a feedforward network hurt performance. Previously, Chelba et al. (2017) had similar findings in language modeling, where using LSTMs truncated to a particular distance improved performance over feedforward networks that were given the same context.

7 Conclusion

In this paper, we investigated the extent to which information provided directly by model structure in classical constituency parsers is still being captured by neural methods. Because neural models function in a substantially different way than classical systems, it could be that they rely on different information when making their decisions. Our findings suggest that, to the contrary, the neural systems are learning to capture many of the same knowledge sources that were previously provided, including the parent-child relations encoded in grammars and the word features induced by lexicons.

Acknowledgments

This work is supported by the DARPA Explainable Artificial Intelligence (XAI) program and the UC Berkeley Savio computational cluster. The second author is supported by an NSF Graduate Research Fellowship.

References

Miguel Ballesteros, Chris Dyer, and Noah A. Smith. 2015. Improved transition-based parsing by modeling characters instead of words with lstms. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 349–359. https://doi.org/10.18653/v1/D15-1041.

Eugene Charniak and Mark Johnson. 2005. Coarse-to-fine n-best parsing and maxent discriminative reranking. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*. Association for Computational Linguistics, pages 173–180. http://www.aclweb.org/anthology/P05-1022.

Kushal Chawla, Sunil Kumar Sahu, and Ashish Anand. 2017. Investigating how well contextual features are captured by bi-directional recurrent neural network models. *arXiv preprint arXiv:1709.00659* .

Ciprian Chelba, Mohammad Norouzi, and Samy Bengio. 2017. N-gram language modeling using recurrent neural network estimation. *arXiv preprint arXiv:1703.10724* .

Do Kook Choe and Eugene Charniak. 2016. Parsing as language modeling. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 2331–2336. https://doi.org/10.18653/v1/D16-1257.

Jan Chorowski, Michał Zapotoczny, and Paweł Rychlikowski. 2016. Read, tag, and parse all at once, or fully-neural dependency parsing. *arXiv preprint arXiv:1609.03441* .

Maximin Coavoux and Benoit Crabbé. 2017. Multilingual lexicalized constituency parsing with word-level auxiliary tasks. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Association for Computational Linguistics, pages 331–336. http://aclweb.org/anthology/E17-2053.

Michael Collins. 1997. Three generative, lexicalised models for statistical parsing. In *35th Annual Meeting of the Association for Computational Linguistics*. http://www.aclweb.org/anthology/P97-1003.

James Cross and Liang Huang. 2016a. Incremental parsing with minimal features using bi-directional lstm. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, pages 32–37. https://doi.org/10.18653/v1/P16-2006.

James Cross and Liang Huang. 2016b. Span-based constituency parsing with a structure-label system and provably optimal dynamic oracles. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 1–11. https://doi.org/10.18653/v1/D16-1001.

Greg Durrett and Dan Klein. 2015. Neural crf parsing. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, pages 302–312. https://doi.org/10.3115/v1/P15-1030.

Chris Dyer, Adhiguna Kuncoro, Miguel Ballesteros, and Noah A. Smith. 2016. Recurrent neural network grammars. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, pages 199–209. https://doi.org/10.18653/v1/N16-1024.

Jenny Rose Finkel, Alex Kleeman, and Christopher D. Manning. 2008. Efficient, feature-based, conditional random field parsing. In *Proceedings of ACL-08: HLT*. Association for Computational Linguistics, Columbus, Ohio, pages 959–967. http://www.aclweb.org/anthology/P08-1109.

David Hall, Greg Durrett, and Dan Klein. 2014. Less grammar, more features. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pages 228–237. https://doi.org/10.3115/v1/P14-1022.

Yoon Kim, Yacine Jernite, David Sontag, and Alexander M Rush. 2016. Character-aware neural language models. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*. AAAI Press, pages 2741–2749.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR* abs/1412.6980. http://arxiv.org/abs/1412.6980.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*. http://www.aclweb.org/anthology/P03-1054.

Terry Koo and Michael Collins. 2010. Efficient third-order dependency parsers. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pages 1–11. http://www.aclweb.org/anthology/P10-1001.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data .

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, pages 260–270. https://doi.org/10.18653/v1/N16-1030.

Percy Liang, Hal Daumé III, and Dan Klein. 2008. Structure compilation: trading structure for features. In *Proceedings of the 25th international conference on Machine learning*. ACM, pages 592–599.

Wang Ling, Chris Dyer, Alan W Black, Isabel Trancoso, Ramon Fermandez, Silvio Amir, Luis Marujo, and Tiago Luis. 2015a. Finding function in form: Compositional character models for open vocabulary word representation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 1520–1530. https://doi.org/10.18653/v1/D15-1176.

Wang Ling, Isabel Trancoso, Chris Dyer, and Alan W Black. 2015b. Character-based neural machine translation. *arXiv preprint arXiv:1511.04586* .

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of lstms to learn syntax-sensitive dependencies. *Transactions of the Association of Computational Linguistics* 4:521–535. http://www.aclweb.org/anthology/Q16-1037.

Jiangming Liu and Yue Zhang. 2017. Shift-reduce constituent parsing with neural lookahead features. *Transactions of the Association of Computational Linguistics* 5:45–58. http://www.aclweb.org/anthology/Q17-1004.

Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of english: The penn treebank. *Comput. Linguist.* 19(2):313–330. http://dl.acm.org/citation.cfm?id=972470.972475.

Graham Neubig, Chris Dyer, Yoav Goldberg, Austin Matthews, Waleed Ammar, Antonios Anastasopoulos, Miguel Ballesteros, David Chiang, Daniel Clothiaux, Trevor Cohn, Kevin Duh, Manaal Faruqui, Cynthia Gan, Dan Garrette, Yangfeng Ji, Lingpeng Kong, Adhiguna Kuncoro, Gaurav Kumar, Chaitanya Malaviya, Paul Michel, Yusuke Oda, Matthew Richardson, Naomi Saphra, Swabha Swayamdipta, and Pengcheng Yin. 2017. Dynet: The dynamic neural network toolkit. *arXiv preprint arXiv:1701.03980* .

Slav Petrov and Dan Klein. 2007. Improved inference for unlexicalized parsing. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*. Association for Computational Linguistics, pages 404–411. http://www.aclweb.org/anthology/N07-1051.

Xing Shi, Inkit Padhi, and Kevin Knight. 2016. Does string-based neural mt learn source syntax? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 1526–1534. https://doi.org/10.18653/v1/D16-1159.

Mitchell Stern, Jacob Andreas, and Dan Klein. 2017. A minimal span-based neural constituency parser. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pages 818–827. https://doi.org/10.18653/v1/P17-1076.

Ben Taskar, Vassil Chatalbashev, Daphne Koller, and Carlos Guestrin. 2005. Learning structured prediction models: A large margin approach. In *Proceedings of the 22nd international conference on Machine learning*. ACM, pages 896–903.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*. Association for Computational Linguistics, Stroudsburg, PA, USA, NAACL '03, pages 173–180. https://doi.org/10.3115/1073445.1073478.

Oriol Vinyals, Łukasz Kaiser, Terry Koo, Slav Petrov, Ilya Sutskever, and Geoffrey Hinton. 2015. Grammar as a foreign language. In *Advances in Neural Information Processing Systems*. pages 2773–2781.

Wenhui Wang and Baobao Chang. 2016. Graph-based dependency parsing with bidirectional lstm. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pages 2306–2315. https://doi.org/10.18653/v1/P16-1218.

Xingxing Zhang, Jianpeng Cheng, and Mirella Lapata. 2017. Dependency parsing as head selection. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Association for Computational Linguistics, pages 665–676. http://www.aclweb.org/anthology/E17-1063.

A Model Hyperparameters and Training Details

Component	Dimensions	Layers
Word Embeddings	100	
Character Embeddings	50	
Character LSTM	100	1
Sentence LSTM	250	2
Label Feedforward Network	250	1

Table 2: The sizes of the components used in our model.

Our model hyperparameters are summarized in Table 2. We train using the Adam optimizer (Kingma and Ba, 2014) with its default hyperparameters for 40 epochs. We evaluate on the development set 4 times per epoch, selecting the model with the highest overall development performance as our final model. When performing a word embedding lookup during training, we randomly replace words by the `<UNK>` token with probability $1/(1 + \text{freq}(w))$, where $\text{freq}(w)$ is the frequency of a word w in the training set. We apply dropout with probability 0.4 before and inside each layer of each LSTM. Our system is implemented in Python using DyNet (Neubig et al., 2017).

B Character LSTM Word Feature Classification

Binary Feature	Majority Class	Char-LSTM Classifier	Binary Feature	Majority Class	Char-LSTM Classifier
all-letters	77.22%	99.77%	suffix = “s”	82.65%	99.99%
has-letter	89.18%	99.97%	suffix = “ed”	92.52%	99.98%
all-lowercase	56.95%	99.95%	suffix = “ing”	93.26%	99.95%
has-lowercase	85.85%	99.90%	suffix = “ion”	97.75%	99.93%
all-uppercase	96.68%	99.90%	suffix = “er”	96.42%	99.97%
has-uppercase	67.77%	99.97%	suffix = “est”	99.63%	99.98%
all-digits	98.38%	99.99%	suffix = “ly”	97.56%	99.99%
has-digit	87.90%	99.91%	suffix = “ity”	99.30%	99.94%
all-punctuation	99.93%	99.98%	suffix = “y”	92.97%	99.93%
has-punctuation	79.04%	99.75%	suffix = “al”	98.48%	99.92%
has-dash	88.89%	99.95%	suffix = “ble”	99.30%	99.90%
has-period	92.55%	99.95%	suffix = “e”	89.57%	99.99%
has-comma	98.02%	99.97%			

Table 3: Classification accuracy for various binary word features using the character LSTM representations for words induced by a pre-trained parser. Performance substantially exceeds that of a majority class classifier in all cases, reaching 99.7% or higher for all features. The majority class is `True` for the first four features in the left column and `False` for the rest.