Abstract

In computational linguistics, it has been shown that hierarchical structures make language models (LMs) more human-like. However, the previous literature has been agnostic about a parsing strategy of the hierarchical models. In this paper, we investigated whether hierarchical structures make LMs more human-like, and if so, which parsing strategy is most cognitively plausible. In order to address this question, we evaluated three LMs against human reading times in Japanese with head-final left-branching structures: Long Short-Term Memory (LSTM) as a sequential model and Recurrent Neural Network Grammars (RNNGs) with top-down and left-corner parsing strategies as hierarchical models. Our computational modeling demonstrated that left-corner RNNGs outperformed top-down RNNGs and LSTM, suggesting that hierarchical and left-corner architectures are more cognitively plausible than top-down or sequential architectures. In addition, the relationships between the cognitive plausibility and (i) perplexity, (ii) parsing, and (iii) beam size will also be discussed.\footnote{Our codes are available at https://github.com/osekilab/RNNG-LC.}

1 Introduction

It has been debated in computational linguistics whether language models (LMs) become more human-like by explicitly modeling hierarchical structures of natural languages. Previous work has revealed that while sequential models such as recurrent neural networks (RNNs) can successfully predict human reading times (Frank and Bod, 2011), there is an advantage in explicitly modeling hierarchical structures (Fossum and Levy, 2012). More recently, RNNs that explicitly model hierarchical structures, namely Recurrent Neural Network Grammars (RNNGs, Dyer et al., 2016), have attracted considerable attention, effectively capturing grammatical dependencies (e.g., subject-verb agreement) much better than RNNs in targeted syntactic evaluations (Kuncoro et al., 2018; Wilcox et al., 2019). In addition, Hale et al. (2018) showed that RNNGs can successfully predict human brain activities, and recommended RNNGs as “a mechanistic model of the syntactic processing that occurs during normal human sentence processing.”

However, this debate has focused almost exclusively on the dichotomy between the hierarchical and sequential models, without reference to the parsing strategies among the hierarchical models. Specifically, although Dyer et al. (2016) and Hale et al. (2018) adopted the vanilla RNNG with a top-down parsing strategy for English with head-initial right-branching structures, Abney and Johnson (1991) and Resnik (1992) suggested that the top-down parsing strategy is not optimal for head-final left-branching structures, and alternatively proposed the left-corner parsing strategy as more human-like parsing strategy.

In this paper, we investigate whether hierarchical structures make LMs more human-like, and if so, which parsing strategy is most cognitively plausible. In order to address this question, we evaluate three LMs against human reading times in Japanese with head-final left-branching structures: Long Short-Term Memory (LSTM) as a sequential model and Recurrent Neural Network Grammars (RNNGs) with top-down and left-corner parsing strategies as hierarchical models.

2 Linking hypothesis

It is well established in psycholinguistics that humans predict next segments during sentence processing, and the less predictable the segment is, the more surprising that segment is. Surprisal theory (Hale, 2001; Levy, 2008) quantifies this predictability of the segment as $-\log p(\text{segment} | \text{context})$, an information-theoretic complexity metric known as surprisal. In line with the previous literature (e.g., Smith and Levy, 2013), we employed this
metric to logarithmically link probabilities estimated from LMs with cognitive efforts measured from humans. Intuitively, the cognitively plausible LMs will compute surprisals with similar trends as human cognitive efforts. Computational models of human sentence processing have been explored by comparing surprisals from various LMs with reading times (e.g., Frank and Bod, 2011) and brain activities (e.g., Frank et al., 2015).

3 Methods

3.1 Language models

Long Short-Term Memory (LSTM): LSTMs are a sequential model that does not model hierarchical structures. We used a 2-layer LSTM with 256 hidden and input dimensions. The implementation by Gulordava et al. (2018) was employed.

Recurrent Neural Network Grammar (RNNG): RNNGs are a hierarchical model that explicitly models hierarchical structures. We experimented with two types of stack-only RNNGs (Kuncoro et al., 2017): top-down RNNGs and left-corner RNNGs (Kuncoro et al., 2018). We used RNNGs that had a 2-layer stack LSTM with 256 hidden and input dimensions. The implementation by Noji and Oseki (2021) was employed. Word-synchronous beam search (Stern et al., 2017) was used for inference. RNNGs were evaluated with six beam sizes $k = \{100, 200, 400, 600, 800, 1000\}$.

3.2 Data sets

Training data: All LMs were trained on the National Institute for Japanese Language and Linguistics Parsed Corpus of Modern Japanese (NPCMJ), that comprises 67,018 sentences annotated with tree structures.$^3$ The sentences were split into subwords by a byte-pair encoding (Sennrich et al., 2016). LSTM used only terminal subwords, while RNNGs used terminal subwords and tree structures, both of which were trained sentence-level for 40 epochs and 3 times with different random seeds.$^5$

Reading time data: All LMs were evaluated against first pass reading times from BCCWJ-EyeTrack (Asahara et al., 2016), that comprises 218 sentences annotated with eye-tracking reading times at each phrasal unit. Following Asahara et al. (2016), the data points (i) not corresponding to the main text or (ii) not fixated were removed. In addition, following Fossum and Levy (2012), the data points that contained subwords “unknown” to the LMs were also removed. Consequently, we included 12,114 data points in the statistical analyses among 19,717 data points in total.

3.3 Evaluation metrics

Psychometric predictive power: We evaluated how well surprisal ($-\log p(\text{segment}|\text{context})$) from each LM could predict human reading times. LMs process the sentences subword-by-subword, while reading times are annotated phrase-by-phrase. Thus, following Wilcox et al. (2020), the surprisal surprisal $I(p)$ was computed as the cumulative sum of surprisals of its constituent subwords $w_1, w_{l+1}, \cdots, w_m$:

$$I(p) = I(w_1, w_{l+1}, \cdots, w_m) = \sum_{i=l}^{m} I(w_i) \quad (1)$$

where $I(w)$ is the surprisal of subword $w$:

$$I(w_i) = -\log p(w_i|w_{<i}) \quad (2)$$

For the statistical analyses, we first trained a baseline regression model with several predictors that are known to affect reading times. Then, we added surprisal estimated from each LM as a predictor and evaluated the decrease in deviance ($\Delta D(LM)$) as psychometric predictive power.$^6$

$$\Delta D(LM) = D_B - D_{LM} \quad (3)$$

where $D_B$ and $D_{LM}$ are deviance of the baseline regression model and the regression model with surprisal, respectively. The details of our regression models are shown in Appendix A.

$^3$ Following Frank and Bod (2011), traces and semantic information were removed in the way described in Manning and Schutze (1999).

In the previous literature, surprisals of the previous segments were also added as a predictor to capture spillover effects (Smith and Levy, 2013). In our experiments, we did not add surprisals of the previous segments because preliminary experiments showed that they were not significant for modeling reading times in all the LMs.
Perplexity and parsing accuracy: Goodkind and Bicknell (2018) demonstrated that perplexity of LMs and their psychometric predictive power are highly correlated. In order to investigate whether this correlation can be observed, perplexities of LMs were calculated based on the sentences in BCCWJ-EyeTrack.

In addition, given that RNNGs also serve as a parser, the correlation between parsing accuracy and psychometric predictive power was also investigated. The evaluation metric of parsing accuracy was the labeled bracketing F1. For this purpose, we used the sentences in NPCMJ because the sentences in BCCWJ-EyeTrack are not annotated with tree structures. Parsing accuracies of RNNGs were calculated based on the tree structures at the top of the final beam in word-synchronous beam search.

4 Results and discussion

The result of our computational modeling is summarized in Figure 1: psychometric predictive power (the vertical axis) is plotted against perplexity (the horizontal axis). In this section, we first analyze psychometric predictive power itself, and then discuss its relationships with (i) perplexity, (ii) parsing, and (iii) beam size.

4.1 Psychometric predictive power

Figure 1 demonstrates that the hierarchical models (top-down/left-corner RNNGs) achieved higher psychometric predictive power than the sequential model (LSTM) and, among the hierarchical models, left-corner RNNGs achieved higher psychometric predictive power than top-down RNNGs. In order to confirm that these differences are statistically meaningful, we performed nested model comparisons. The result of nested model comparisons is summarized in Table 1, where the best result from each LM was compared. The significance threshold at $\alpha = 0.0056$ was imposed by the Bonferroni correction motivated by 9 tests ($0.05/9$).

First, nested model comparisons revealed that psychometric predictive power was significant for all LMs relative to the baseline regression model. The point here is that surprisals computed by LMs do explain human reading times in Japanese, generalizing the previous results in English.

Second, the hierarchical models (top-down/left-corner RNNGs) significantly outperformed the sequential model (LSTM), and the sequential model did not account for unique variances that the hierarchical models cannot explain. This result aligns with the previous observation that LMs become more human-like by explicitly modeling hierarchical structures on top of linear strings (Kuncoro et al., 2018; Wilcox et al., 2019; Hale et al., 2018).

Finally, among the hierarchical models, left-corner RNNGs significantly outperformed top-down RNNGs, and top-down RNNGs did not account for unique variances that left-corner RNNGs cannot explain. This result corroborates Abney and Johnson (1991) from an information-theoretic perspective: the left-corner parsing strategy is more cognitively plausible than the top-down and bottom-up parsing strategies.

Here we can conclude from these results that LMs become more human-like by explicitly modeling hierarchical structures and, most importantly, the left-corner parsing strategy was more cognitively plausible than the top-down parsing strategy.

4.2 Perplexity

In this subsection, we discuss the relationship between perplexity and psychometric predictive power. First, Figure 1 indicates that, among the hierarchical models, left-corner RNNGs, which achieved higher psychometric predictive power, also achieved lower perplexity than top-down RNNGs. Overall, the correlation between perplexity and psychometric predictive power of the hierarchical models was robust: the lower perplexity RNNGs have, the higher psychometric predictive power.
Table 1: The result of nested model comparisons with the best result from each LM. TD and LC indicate top-down and left-corner RNNGs, respectively. The significance threshold at $\alpha = 0.0056$ was imposed by the Bonferroni correction motivated by 9 tests ($0.05/9$).

<table>
<thead>
<tr>
<th>Model Pair</th>
<th>$\chi^2$</th>
<th>df</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline &lt; LSTM</td>
<td>260.26</td>
<td>1</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Baseline &lt; TD</td>
<td>299.5</td>
<td>1</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Baseline &lt; LC</td>
<td>306.78</td>
<td>1</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>LSTM &lt; TD</td>
<td>42.183</td>
<td>1</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>LSTM &lt; LC</td>
<td>47.229</td>
<td>1</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>TD &lt; LSTM</td>
<td>2.9406</td>
<td>1</td>
<td>0.08638</td>
</tr>
<tr>
<td>TD &lt; LC</td>
<td>11.84</td>
<td>1</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>LC &lt; LSTM</td>
<td>0.708</td>
<td>1</td>
<td>0.4001</td>
</tr>
<tr>
<td>LC &lt; TD</td>
<td>4.5609</td>
<td>1</td>
<td>0.03271</td>
</tr>
</tbody>
</table>

In sharp contrast, the correlation did not hold for the sequential model, where LSTMs achieved better perplexity, but worse psychometric predictive power than RNNGs with similar or even worse perplexity, corroborating Goodkind and Bicknell (2018) that LSTM stands out as an outlier of the correlation between perplexity and psychometric predictive power.

Kuribayashi et al. (2021) recently showed that the correlation between perplexity and psychometric predictive power cannot be generalized to Japanese. They proposed that LMs are trained to flatten information density and thus satisfy the Uniform Information Density (UID) assumption (Genzel and Charniak, 2002; Levy, 2005; Jaeger and Levy, 2007), but information density in Japanese turned out not to be empirically uniform and far from the idealized UID assumption. At first, this proposal appears to be inconsistent with our results, but notice that Kuribayashi et al. (2021) only tested sequential models. Here we would like to suggest that, unlike sequential models, hierarchical models can be trained to be human-like, even in languages far from the idealized UID assumption.

### 4.4 Beam size

Finally, we discuss the relationship between beam size and psychometric predictive power. The important generalization here is that, although top-down RNNGs improved in psychometric predictive power, perplexity, and parsing accuracy only when the beam size increased, left-corner RNNGs consistently performed well even with a small beam size. We interpret this generalization as demonstrating that left-corner RNNGs may be more human-like than top-down RNNGs in that they can infer the correct tree structure even with a small beam size comparable to humans. In order to reinforce this reasoning, we discuss (i) why left-corner RNNGs can infer the correct tree structure with a small beam size, and (ii) why inference with a small beam size is comparable to humans.

First, we show why left-corner RNNGs can infer the correct tree structure with a small beam size.

Second, we show why inference with a small beam size is comparable to humans. In fact, Jurafsky (1996) suggested that full parallel processing...
and serial processing are not appropriate for human sentence processing, and instead partial parallel processing via beam search that prunes low probability structures is cognitively plausible. From this perspective, we may argue that inference with a small beam size is comparable to humans, and left-corner RNNGs, which can infer the correct tree structure with a small beam size, may be more human-like than top-down RNNGs.

As an anonymous reviewer correctly pointed out, however, given that Jurafsky (1996) proposed beam search that only keeps structures with a probability within a multiple of 3.8 to 5.6 of the probability of the most probable structure, the number of structures within such a relative beam could be extremely large, especially if the probabilities of the structures within the beam are similar. In order to address this point, we computed an empirical beam size comparable to humans. Specifically, we calculated the number of structures with a probability more than $1/3.8$ and $1/5.6$ of the probability of the most probable structure, among structures within the word beam which corresponds to the beam defined in Jurafsky (1996). The results showed that, even with the largest word beam size of 100, the average number of structures through derivations within the proposed relative beam turned out to be empirically small: between 3.05 and 4.14 for top-down RNNGs and between 4.08 and 5.68 for left-corner RNNGs. The details of the results are shown in Appendix C. We believe that these results do not affect our argument that inference with a small beam size is comparable to humans.

These discussions taken together, we could still argue that left-corner RNNGs, which can infer the correct tree structure with a small beam size, may be more human-like than top-down RNNGs. In addition, given that larger beam sizes make LMs more computationally expensive, these results also suggest that left-corner RNNGs are more efficient.

5 Limitations and future work

Interestingly, Wilcox et al. (2020) demonstrated that top-down RNNGs underperformed LSTMs in predicting human reading times in English, which appears to be contradictory to our results in Japanese. We would like to suggest that this discrepancy can be attributed to the difference in the languages tested in these papers. In fact, Kuribayashi et al. (2021) have shown that several established results in English cannot be straightforwardly generalized to Japanese.

In addition, Wilcox et al. (2020) found that n-gram language models outperformed various neural language models, while Merkx and Frank (2021) observed that Transformers (Vaswani et al., 2017) outperformed LSTMs in modeling self-paced reading times and N400 brain activities, but not in predicting eye-tracking reading times.

In order to address these limitations, we plan to conduct detailed comparisons between English (Dundee Corpus, Kennedy et al., 2003) and Japanese (BCCWJ-EyeTrack, Asahara et al., 2016) with RNNGs, incorporating n-gram language models and Transformers as baselines in future work.

6 Conclusion

In this paper, we investigated whether hierarchical structures make LMs more human-like, and if so, which parsing strategy is most cognitively plausible. Our computational modeling demonstrated that left-corner RNNGs outperformed top-down RNNGs and LSTM, suggesting that hierarchical and left corner architectures are more cognitively plausible than top-down or sequential architectures. Moreover, lower perplexities and higher parsing accuracies of the hierarchical models were strongly correlated with the higher psychometric predictive power, but the correlation did not hold for the sequential model. In addition, left-corner RNNGs may be more human-like than top-down RNNGs in that they can infer the correct tree structure with a small beam size comparable to humans.

Acknowledgements

We would like to thank three anonymous reviewers for their insightful suggestions. This work was supported by JSPS KAKENHI Grant Number 19H04990 and the National Institute for Japanese Language and Linguistics (NINJAL) Collaborative Research Project “Computational Psycholinguistics of Language Processing with Large Corpora.”

References


Masayuki Asahara, Kikuo Maekawa, Mizuho Imada, Sachi Kato, and Hikari Konishi. 2014. Archiving and analysing techniques of the ultra-large-scale


The logarithmic reading time ($\log(\text{RT})$) was modeled using the following linear mixed-effects model as a baseline regression model:

$$
\log(\text{RT}) \sim \text{length} + \text{prev_length} + \text{freq} + \text{prev_freq} + \text{is_first} + \text{is_last} + \text{is_second_last} + \text{screenN} + \text{lineN} + \text{segmentN} + (1|\text{article}) + (1|\text{subj}).
$$

(A) Details of our regression model

The logarithmic reading time ($\log(\text{RT})$) was modeled using the following linear mixed-effects model as a baseline regression model:

$$
\log(\text{RT}) \sim \text{length} + \text{prev_length} + \text{freq} + \text{prev_freq} + \text{is_first} + \text{is_last} + \text{is_second_last} + \text{screenN} + \text{lineN} + \text{segmentN} + (1|\text{article}) + (1|\text{subj}).
$$

(B) Relationship between parsing accuracy and psychometric predictive power

The relationship between parsing accuracy and psychometric predictive power is summarized in Figure 2: psychometric predictive power is plotted against parsing accuracy (F1). Just like perplexity, left-corner RNNGs, which achieved higher psychometric predictive power, also achieved the higher parsing accuracy than top-down RNNGs. Overall, the correlation between parsing accuracy and psychometric predictive power of the hierarchical models was robust: the higher parsing accuracy RNNGs have, the higher psychometric predictive power they also have.

(C) Empirical beam size in Jurafsky (1996)

Table 3 shows the average number of structures through derivations within the beam proposed in Jurafsky (1996). Specifically, we calculated the number of structures with a probability more than 1/3.8 and 1/5.6 of the probability of the most probable structure, among structures within the word beam which corresponds to the beam defined in Jurafsky (1996). The average number of structures within the proposed relative beam turned out to be empirically small.
D Hyparameters and available codes

Table 4 shows the hyperparameters of LMs. The code of LSTM we employed (Gulordava et al., 2018) is available at https://github.com/facebookresearch/colorlessgreenRNNs. Surprisals of LSTM were calculated with the code implemented by van Schijndel and Linzen (2018), which is available at https://github.com/vansky/neural-complexity. The code of RNNGs we employed (Noji and Oseki, 2021) is also available at https://github.com/aistairc/rnng-pytorch. We used a byte-pair encoding implemented in sentencepiece (Kudo and Richardson, 2018), which is available at https://github.com/google/sentencepiece. We set character coverage to 0.9995, and vocabulary size to 8,000.

E Dataset split ratio

Sentences in the training data, NPCMJ, are from 14 sources. We used 90% of sentences in each source as a training set, and 5% of sentences as a validation set. The remaining 5% were used as a test set to calculate paring accuracies of RNNGs in Section 4 and perplexities of LMs in Appendix F.

F Relationship between perplexity on the NPCMJ test set and psychometric predictive power

We additionally investigated the relationship between perplexity calculated based on the sentences in the NPCMJ test set and psychometric predictive power. The result is shown in Figure 3: psychometric predictive power is plotted against perplexity on the NPCMJ test set. Although the perplexities of all LMs were overall lower, there was no substantial difference with the result shown in Figure 1. The difference in the corpus domain may cause the overall difference in perplexity.
<table>
<thead>
<tr>
<th>Name</th>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>length</td>
<td>int</td>
<td>Number of characters in the segment</td>
</tr>
<tr>
<td>prev_length</td>
<td>int</td>
<td>Number of characters in the previous segment</td>
</tr>
<tr>
<td>freq</td>
<td>num</td>
<td>Logarithm of the geometric mean of the word frequencies in the segment</td>
</tr>
<tr>
<td>prev_freq</td>
<td>num</td>
<td>Logarithm of the geometric mean of the word frequencies in the previous segment</td>
</tr>
<tr>
<td>is_first</td>
<td>factor</td>
<td>Whether the segment is the first on a line</td>
</tr>
<tr>
<td>is_last</td>
<td>factor</td>
<td>Whether the segment is the last on a line</td>
</tr>
<tr>
<td>is_second_last</td>
<td>factor</td>
<td>Whether the segment is the second to last on a line</td>
</tr>
<tr>
<td>screenN</td>
<td>int</td>
<td>Screen display order</td>
</tr>
<tr>
<td>lineN</td>
<td>int</td>
<td>Line display order</td>
</tr>
<tr>
<td>segmentN</td>
<td>int</td>
<td>Segment display order</td>
</tr>
<tr>
<td>article</td>
<td>factor</td>
<td>Article ID</td>
</tr>
<tr>
<td>subj</td>
<td>factor</td>
<td>Participant ID</td>
</tr>
</tbody>
</table>

Table 2: Descriptions for the factors used in our experiments.

<table>
<thead>
<tr>
<th>Model</th>
<th>Threshold</th>
<th>Word beam size</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>10</td>
<td>20</td>
</tr>
<tr>
<td>TD</td>
<td>1/3.8</td>
<td>1.67 (±0.0)</td>
</tr>
<tr>
<td></td>
<td>1/5.6</td>
<td>1.95 (±0.0)</td>
</tr>
<tr>
<td>LC</td>
<td>1/3.8</td>
<td>2.58 (±0.1)</td>
</tr>
<tr>
<td></td>
<td>1/5.6</td>
<td>3.14 (±0.1)</td>
</tr>
</tbody>
</table>

Table 3: The average number of structures with a probability more than 1/3.8 and 1/5.6 of the probability of the most probable structure, among structures within the word beam which corresponds to the beam defined in Jurafsky (1996). TD and LC indicate top-down and left-corner RNNGs, respectively. Average scores with standard deviations across different random seeds are reported.

<table>
<thead>
<tr>
<th></th>
<th>optimizer</th>
<th>learning rate</th>
<th>dropout</th>
<th>batch size</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSTM</td>
<td>SGD</td>
<td>20</td>
<td>0.2</td>
<td>64</td>
</tr>
<tr>
<td>Top-down RNNGs</td>
<td>adam</td>
<td>0.001</td>
<td>0.3</td>
<td>64</td>
</tr>
<tr>
<td>Left-corner RNNGs</td>
<td>adam</td>
<td>0.001</td>
<td>0.3</td>
<td>64</td>
</tr>
</tbody>
</table>

Table 4: Hyperparameters of LMs.