Can language models learn constraints on gap-filler dependency? Case of Japanese relative clause islands

Maho Takahashi
Department of Linguistics
University of California, San Diego
mtakahas@ucsd.edu

1 Introduction

Several recent studies have evaluated whether neural language models (LMs) such as the one with the long short-term memory (LSTM) architecture can acquire various syntactic phenomena (Linzen et al. 2016; Bernardy & Lappin 2017; Kuncoro et al. 2018; Gulordava et al. 2018; Futrell et al. 2018; Marvin & Linzen 2018; Wilcox et al. 2018, 2021). These studies show in particular that LMs can learn English long-distance extractions such as *wh*-questions (=1a), as well as the constraint that they cannot be done across certain structures known as *islands* (=1b, with the bracketed relative clause acting as an island).

(1) a. **What**$_i$ did Rebecca believe you claimed that the professor discussed __$_i$?
b. ***What**$_i$ did Rebecca believe your claim [that the professor discussed __$_i$]?

Target language of those studies has largely been limited to English, where most dependencies have the filler-gap order. Yet given the proposal that islands are universal constraints (Ross, 1967), they are predicted to exist regardless of the linear order between a filler and a gap. This study thus focuses on Japanese, where operations such as relativization result in gap-filler dependency, given its head-final property. As relative clauses constitute an island in Japanese (Saito 1985), relativization out of another relative clause should be considered ill-formed. While that is the case in English (=2a), such a construction has been deemed possible in Japanese (=2b) (Kuno 1973; Ishizuka 2009).

(2) a. *English – filler-gap*
*This is **the novel**$_j$ that [the professor$_i$ that [__$_i$ wrote __$_j$]] is very proud.

b. *Japanese – gap-filler*
?This is [[__$_i$ __$_j$ wrote] the professor$_i$ is very proud] **the novel**$_j$.

Contrary to the previous judgments, however, a formal acceptability experiment in Takahashi and Goodall (to appear) revealed that Japanese speakers are in fact sensitive to the violation of a relative clause island. In light of the evidence that filler-gap and gap-filler dependencies both exhibit some amount of sensitivity to relative clause islands, a remaining question is whether Japanese LMs are capable of learning the constraints on gap-filler dependency, on a par with humans. In the current study, I attempt to answer this question by evaluating the performance of a couple of pre-trained Japanese LMs on a set of sentences, some of which involve a relative clause island violation.

2 Experiment

2.1 Models

Two variations of LMs pre-trained with Japanese texts were included in this evaluation. First, I tested a LSTM trained on news articles and the Japanese part of Wikipedia by Kuribayashi et al. (2021)[1]. The training dataset consist of approximately 5M sentences (146M subword units, which were divided by byte-pair encoding), and the training involved 1K parameter updates. Second, we tested a Transformer (GPT-2) trained by Colorful Scoop[2],

[1] https://github.com/kuribayashi4/surprisal_reading_time_en_ja

[2] https://huggingface.co/colorfulscoop/gpt2-small-ja

Proceedings of the Society for Computation in Linguistics (SCiL) 2023, pages 357-360.
Amherst, Massachusetts, June 15-17, 2023

(3) a. *No-extraction baseline*

pro$_i$	*kai-ta*	*hon-ga*	*news-de*	*tokusyuu-sa-re-ta*	*node,*
	write-PST	book-NOM	news-on	feature-do-PASS-PST	because
kyoujyu$_i$*-wa*	*tokuige-da.*				
professor-TOP	proud-COP				

'Because the book that (he$_i$) wrote was featured in the news, **the professor$_i$** looks proud.'

b. *Extraction out of a non-island*

[__$_i$	*hon-o*	*kai-ta*	*koto]-ga*	*news-de*	*tokusyuu-sa-re-ta*
	book-ACC	write-PST	fact-NOM	news-on	feature-do-PASS-PST
kyoujyu$_i$*-wa*	*tokuige-da.*				
professor-TOP	proud-COP				

'**The professor$_i$** who [the fact that __$_i$ wrote a book was featured in the news] looks proud.'

c. *Extraction out of an island*

[[__$_i$ __$_j$	*kai-ta]*	*hon$_j$-ga*	*news-de*	*tokusyuu-sa-re-ta]*
	write-PST	book-NOM	fact-NOM	feature-do-PASS-PST
kyoujyu$_i$*-wa*	*tokuige-da.*			
professor-TOP	proud-COP			

'**The professor$_i$** who [the book$_j$ that [__$_i$ wrote __$_j$] was featured in the news] looks proud.'

which was also trained on the Japanese Wikipedia dataset (token size: 540M, 110M parameters).

2.2 Data

The models were presented with three types of sentences exemplified in (3). *The professor* in (3a) is base-generated while its corresponding null pronoun (*pro*) occupies the gap position. As (3a) does not involve extraction, the dependency between *the professor* and its corresponding *pro* is not bound by islands. (3b) does involve extraction of *the professor*, but out of a complex noun phrase headed by *koto* 'the fact', which has been claimed to be not an island in Japanese (Omaki et al. 2020). (3c) exemplifies a structure with extraction out of the relative clause island.

2.3 Procedure

In line with the previous studies evaluating the grammatical knowledge of LMs, surprisal (Hale 2001; Levy 2008) was calculated for each word $S(w_k)$ upon seeing the word w_k given h_{k-1}, the hidden state after processing all the previous words in a sentence: $-\log_2 \mathbb{P}(w_k|h_{k-1})$. I compared the mean surprisal of the gray region in (3), which includes the noun phrase (*kyoujyu* 'professor') that has been extracted out of an embedded clause in (3b,c). I predicted that the mean surprisal value for the region would be (3a) < (3b) < (3c) if the LMs have learned both long-distance extraction (which lowers acceptability in and of itself; Fodor 1978) and the relative clause island constraint; in contrast, mean surprisal value would be (3a) < (3b) = (3c) if the LMs have learned only the former.

2.4 Results

As Table 1 shows, mean surprisal values produced by the LSTM model did not differ across sentence types (3a-c), as confirmed by a series of one-sample *t*-tests (all above p=0.05). In contrast, GPT-2's mean surprisal for sentences of type (3a) was significantly smaller than the one of both (3b) (t=-4.32, p<0.001) and (3c) (t=4.54, p<0.001), in accord with the fact that the former does not involve extraction but the latter do. Critically, however, the mean surprisal values of (3b) and (3c) were not significantly different (t=0.52, p=0.61), contrary to the experimental result with human participants (Takahashi & Goodall to appear).

	no_ext (3a)	ext_noisl (3b)	ext_isl (3c)
LSTM	4.63	4.86	4.82
GPT-2	7.15	9.21	9.13

Table 1: Mean surprisal values of the critical regions (greyed in (3)) produced by LSTM & Transformer models.

3 Discussion

The results reported here suggest that only some of the Japanese LMs (namely, GPT-2)

distinguish between sentences with and without a long-distance extraction, and they do not seem to show sensitivity to the relative clause island. One might argue that the poor performance of Japanese LSTM LM is due to the small size of training data. In Wilcox et al. (2021), however, LSTM models trained with approximately 90M tokens demonstrated the knowledge of English islands. It is therefore not unreasonable to expect that the Japanese LSTM would learn island constraints with 146M-subword training data, given the previous success with English.

Our findings have several implications: First, they suggest that Transformer may be a better architecture for capturing gap-filler dependencies than LSTM for Japanese, which goes against the observed pattern among English LMs that RNNs may model linguistic competence better than the Transformers (Wilcox et al. 2021). Second, the fact that neither LSTM nor GPT-2 showed sensitivity to relative clause islands contradicts the experimental evidence provided in Takahashi and Goodall (to appear), as well as the island sensitivity exhibited by LMs reported in previous studies. Recall that a relative clause island violation in Japanese has been considered possible (Kuno 1973; Ishizuka 2009). Island effects that can only be revealed through formal acceptability experiments (like the one conducted by Takahashi and Goodall) have been reported and are known as *subliminal* islands (Almeida 2014; Keshev & Meltzer-Asscher 2019). Further investigation of whether such subliminal island effects are learnable by LMs, using a broader range of structures, will contribute to the ongoing discussion about the nature and gradience in island effects across languages.

Acknowledgements

I would like to thank two anonymous reviewers as well as Leon Bergen for their comments and feedback.

References

Almeida, D. (2014). Subliminal wh-islands in brazilian Portuguese and the consequenceS for Syntactic theory. *Revista Da ABRALIN*, 13(2), 55–93.

Bernardy, J. P., & Lappin, S. (2017). Using deep neural networks to learn syntactic agreement. *Linguistic Issues in Language Technology*, 15(2), 1–15.

Fodor, J. D. (1978). Parsing strategies and constraints on transformations. *Linguistic Inquiry*, 9(3), 427–473.

Futrell, R., Wilcox, E., Morita, T., and Levy, R. (2018). Rnns as psycholinguistic subjects: Syntactic state and grammatical dependency. *arXiv:1809.01329*.

Gulordava, K., Bojanowski, P., Grave, E., Linzen, T., & Baroni, M. (2018). Colorless green recurrent networks dream hierarchically. *arXiv:1803.11138*.

Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, 1–8.

Ishizuka, T. (2009). CNPC Violations and Possessor Raising in Japanese. In Potter, D., & Storoshenko, D. R. (Eds.), *Proceedings of the 2nd International Conference on East Asian Linguistics* (Issue 1985).

Keshev, M., & Meltzer-Asscher, A. (2019). A processing-based account of subliminal wh-island effects. *Natural Language and Linguistic Theory*, 37(2), 621–657.

Kuncoro, A., Dyer, C., Hale, J., Yogatama, D., Clark, S., & Blunsom, P. (2018). LSTMs can learn syntax-sensitive dependencies well, but modeling structure makes them better. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics,* 1426–1436.

Kuno, S. 1973. *The structure of the Japanese language*. Cambridge, Mass.: MIT Press.

Kuribayashi, T., Oseki, Y., Ito, T., Yoshida, R., Asahara, M., & Inui, K. (2021). Lower perplexity is not always human-like. *Proceedings of ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, 106(3), 1126–1177.

Linzen, T., Dupoux, E., & Goldberg, Y. (2016). Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4, 521–535.

Marvin, R., & Linzen, T. (2018). Targeted syntactic evaluation of language models. *arXiv:1808.09031*.

Omaki, A., Fukuda, S., Nakao, C., & Polinsky, M. (2020). Subextraction in Japanese and subject-object symmetry. *Natural Language & Linguistic Theory*, 38, 627–669.

Ross, J. R. (1967). *Constraints on variables in syntax*.

Saito, M. (1985). *Some asymmetries in Japanese and their theoretical implications* (Doctoral dissertation, MIT).

Takahashi, M. and Goodall, G (to appear). Island sensitivity with relativization in Japanese: The case of double relatives. *Proceedings from the 57th annual meeting of the Chicago Linguistic Society*.

Wilcox, E., Levy, R., Morita, T., & Futrell, R. (2018). What do RNN language models learn about filler-gap dependencies?. *arXiv:1809.00042*.

Wilcox, E. G., Futrell, R., & Levy, R. (2021). *Using Computational Models to Test Syntactic Learnability*.