Expectation Adaptation Models Hindi Preverbal Constituent Ordering

Sidharth Ranjan[1] Rajakrishnan Rajkumar[2] Sumeet Agarwal[1]
IIT Delhi[1], IISER Bhopal[2]
sidharth.ranjan03@gmail.com, rajak@iiserb.ac.in, sumeet@iitd.ac.in

In this study, we investigate if adapting a neural language model's expectation helps it to predict preverbal constituent ordering in Hindi, a predominantly SOV language with flexible word order. Prior work in Hindi has shown that scrambling of word-order is affected by information structure constraints in discourse. Going beyond, we investigate their processing mechanisms by performing three distinct experiments *viz.*, *genre adaptation, syntactic adaptation, and lexical adaptation* using adaptive LSTM surprisal models for the task of word-order prediction. In a recent study, van Schijndel and Linzen (2018) showed that adaptive LSTM language models (LMs) significantly improve the ability to predict human reading times over non-adaptive surprisal. Furthermore, Prasad et al. (2019) demonstrated that neural LMs track abstract properties of the sentences where learned representations can be organized in a linguistically interpretable manner.

First, we set up a framework to generate meaning-equivalent grammatical variants corresponding to sentences in the Hindi-Urdu Treebank corpus (Bhatt et al. 2009; HUTB) of written text from news-wire domain by permuting their preverbal constituents. The first sentence (Example 1 in Table 1) is the reference sentence obtained from the HUTB corpus and considered the most preferred choice despite other legitimate grammatical variant (Example 2-3 in Table 1) that could have been possibly produced. Subsequently, we deployed a logistic regression model to predict HUTB reference sentences (amidst variants expressing the same meaning) using various cognitively motivated sentence-level features, *viz.,* dependency length, surprisal, and adaptive surprisal (van Schijndel and Linzen 2018). Corresponding to the aforementioned experiments, we estimated *genre-based adaptive LSTM surprisal* on our test dataset by adapting the base LSTM LM to the Hindi news-wire section of CIIL corpus (Ramamoorthy et al. 2019), *syntactically adapted LSTM surprisal* by adapting base LSTM LM to only object-fronted sentences from CIIL corpus, and, finally, *lexically adapted LSTM surprisal* by adapting base LSTM LM to only one preceding sentence of each test sentence in the HUTB corpus. Pearson's correlation coefficients revealed that adaptive LSTM surprisal is highly correlated with all other surprisal measures (in particular base LSTM surprisal).

Corroborating the previous findings, our results evinced the effects of adaptation in word-order preferences in Hindi. Regression results pertaining to genre adaptation (see Table 2) suggest that news genre adapted LSTM LM learned the abstract properties therein *viz.,* occurrence of temporal information and pronoun early in the sentence, etc. Further analyses revealed successful adaptation across different constructions, *viz.,* active and passive voice sentences, and conjunct verbs. However, adaptation was not effective for certain non-canonical constructions. The subsequent two experiments focus on these non-canonical constructions where we investigate if adapting to their syntax (*syntactic adaptation*) or preceding sentence in the discourse (*lexical adaptation*) predicts the non-canonical syntactic choices. The syntactic adaptation results suggest that adaptation is sensitive to syntactic structures relating to direct-object fronted and indirect-object fronted sentences (see Table 3). At last, for lexical adaptation experiment, we incorporate an information status score (see Table 1 for illustration) reflecting *given* vs *new* considerations as a control into our regression model. Languages hold fast *give*-before-*new* principle by assessing elements already salient in the discourse (by previous mention) prior to the new content. The results (see Table 4) evinced that previous sentence adaptation accounts for information structure beyond *Given-New* ordering preferences in predicting reference sentences having fronted indirect objects over variants containing

the canonical word-order. Overall, we conclude that adaptation captures not only the stylistic patterns and syntactic structures but also discourse effects to some extent. In future inquiries, we plan to investigate further, the exact nature of information structure constraints captured by adaptive LSTM surprisal in lexical adaptation experiment.

Previous sentence:
amar ujala-ki bhumika nispach rehti hai
Amar Ujala-GEN role fair remain be.PRS.SG
Amar Ujala's role remains fair.

Reference sentence:
1. amar ujala-ko **yah** sukravar-ko daak-se prapt hua (Given-Given)
Amar Ujala-ACC it friday-on post-INST receive be.PST.SG
Amar Ujala received it by post on Friday.

Variant sentence:
2. **yah** *amar ujala ko sukravar ko daak se prapt hua* (Given-Given)
3. *sukravar ko* **yah** *amar ujala ko daak se prapt hua* (New-Given)

Score: Given-New = 1; New-Given = -1; New-New/Given-Given = 0
Given tag assigned to subject/object constituents if it has pronoun as head or previous mention content words else *New tag* was assigned

Table 1: Information Status Score Illustration

Predictor	Coefficient	Std. Error	z-value
intercept	0.02	0.013	1.86
trigram surprisal	-2.24***	0.041	-54.97
dependency length	-0.23***	0.018	-12.7
pcfg surprisal	-0.97***	0.03	-32.09
lstm surprisal	-1.92***	0.057	-33.22
adaptive lstm surprisal	**-3.20*** **	0.062	-51.69

Table 2: Genre adaptation regression results (158891 data points)

Predictor	Coefficient	Std. Error	z-value
intercept	0.08	0.093	0.883
trigram surprisal	-0.65**	0.235	-2.763
dependency length	-0.37*	0.160	-2.310
pcfg surprisal	0.18	0.159	1.141
lstm surprisal	2.59***	0.347	7.497
adaptive lstm surprisal	**-6.74*** **	0.453	-14.869

(a) Direct objects (DO; 1663 points) fronted adaptation

Predictor	Coefficient	Std. Error	z-value
intercept	-0.01	0.131	-0.072
trigram surprisal	-2.19***	0.356	-6.157
dependency length	-0.11	0.138	-0.782
pcfg surprisal	-1.39***	0.278	-5.028
lstm surprisal	3.48***	0.532	6.541
adaptive lstm surprisal	**-6.76*** **	0.644	-10.494

(b) Indirect objects (IO; 1353 points) fronted adaptation

Table 3: Syntactic adaptation regression results

Predictor	Coefficient	Std. Error	z-value
intercept	0.005	0.073	0.069
trigram surprisal	-1.79***	0.251	-7.146
dependency length	-0.59***	0.092	-6.450
pcfg surprisal	-0.13	0.153	-0.853
info status score	0.28***	0.055	5.044
lstm surprisal	0.12	0.837	0.148
adaptive lstm surprisal	-1.59	0.85	-1.872

(a) Direct objects (DO; 1663 points) fronted adaptation

Predictor	Coefficient	Std. Error	z-value
intercept	0.14	0.115	1.241
trigram surprisal	-2.62***	0.405	-6.475
dependency length	-0.10	0.156	-0.656
pcfg surprisal	-1.69***	0.285	-5.949
info status score	0.04	0.080	0.433
lstm surprisal	11.61***	2.745	4.229
adaptive lstm surprisal	**-14.28*** **	2.807	-5.089

(b) Indirect objects (IO; 1353 points) fronted adaptation

Table 4: Lexical adaptation regression results

References

Bhatt, R., Narasimhan, B., Palmer, M., Rambow, O., Sharma, D. M., and Xia, F. (2009). A multi-representational and multi-layered treebank for Hindi/urdu. In *Proceedings of the Third Linguistic Annotation Workshop*, ACL-IJCNLP '09, pages 186–189, Stroudsburg, PA, USA. Association for Computational Linguistics.

Prasad, G., van Schijndel, M., and Linzen, T. (2019). Using priming to uncover the organization of syntactic representations in neural language models. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 66–76.

Ramamoorthy, L., Choudhary, N., Singh, J., Richa, Sinha, A., Mishra, D. K., Tripathi, A. K., Debsharma, A., Awasthi, S. K., and Pathak, M. (2019). *A Gold Standard Hindi Raw Text Corpus*. Central Institute of Indian Languages (CIIL), Mysore.

van Schijndel, M. and Linzen, T. (2018). A neural model of adaptation in reading. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4704–4710.

[0]We thank Prof. Marten van Schijndel and two anonymous reviewers for the insightful comments and feedback.