Transformer Performance on Case in Balto-Slavic Languages

Lorenss Martinsons
Yale University
lorenss.martinsons@yale.edu

1 Introduction

Advances in natural language processing (NLP) have sparked research in language models' performance in many English grammatical phenomena such as subject-verb agreement (Gulordava et al., Linzen et al., Hu et al., Marvin and Linzen) filler-gap dependencies (Hu et al., Wilcox et al., Wilcox et al., Kobzeva et al.), negative polarity items (Hu et al., Marvin and Linzen, Warstadt et al.), among many others. However, there exists a concerning gap in research rigorously evaluating non-English language models' capabilities (Nicholas and Bhatia), and especially, in multilingual language models (Micallef et al., OpenAI). My research aims to shed light on four major yet understudied Balto-Slavic languages - Russain, Ukrainian, Lithuanian, and Latvian. These languages use different scripts, limiting the knowledge that may come from shared tokens. Additionally, they provide a diverse look into the development of modern NLP technology in currently geopolitically important parts of the world.

In Balto-Slavic languages, adjectives, pronouns, and verbs must agree in gender, number, and case with the nouns they modify or are associated with, often bounded by long-range syntactic dependencies. As such, in real-world applications, correct case use is vital to the performance of language models. Currently, however, research on case agreement in general is very limited (Edmiston, Rochereau et al., Ravfogel et al., Ravfogel et al.), and though there is some NLP research in Balto-Slavic languages, there is no contemporary work on case agreement or the performance of multilingual models. To investigate understanding of case, we employ a special context in which genitive case is not tied to a governing predicate (verb/preposition) – case assignment from negation. For instance, the Russian sentence "у меня есть *книга*" (I have a book) negates to "у меня нет *книги*" (I don't have a book), where the nominative книга (book) assumes the genitive form книги. This property holds for most eastern Balto-Slavic languages. The representation of negation as an operator has broad downstream consequences in various NLP tasks (Morante and Blanco), and, as such, is a robust heuristic for language models' performance.

У меня **нет**,	как решила мама,	чая
I don't have	*as mom decided*	*tea*
cue	interjection	target

Figure 1: The cue (in bold) assigns case to the target. Meanwhile, the noun in the interjection (underlined) acts as an attractor.

2 Methods

We construct pairs of positive/negative sentences that induce either the nominative or genitive case in a target constituent (Figure 1).

Cue Two test conditions, **NOM** and **GEN**, differing by the case they assign to the target.

Context Syntactically unintegrated inserts (here called *interjections*) are useful in establishing a syntax-sensitive context that is not solvable by simple adjacency heuristics. We systematically vary the syntactic interjections, by noun choices, length, and attractor location.

Target The target noun takes two forms, in agreement with the cue (**NOM**/ **GEN**). A robust comparison required tight constraints: we make sure sub-word tokenization sets up robust comparisons for the words across all the models, while assuming a probability floor to help ensure the intended target was predicted, and avoid case conflation between singular genitive and plural nominative[1][2].

[1] A common pattern in Balto-Slavic languages

[2] In total, for all languages, we generated 9760 total test cases.

Proceedings of the Society for Computation in Linguistics (SCiL) 2024, pages 285-288.
Irvine, California, June 27-29, 2024

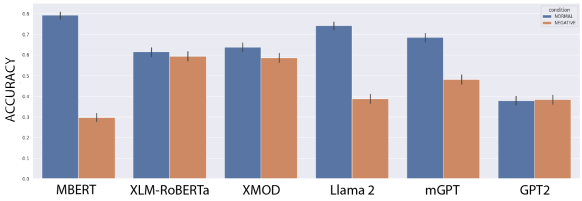

Figure 2: Accuracy by condition (blue: **NOM**, orange: **GEN**)

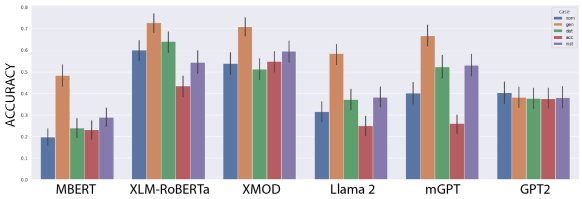

Figure 3: Accuracy by case in **GEN** conditions

I tested 6 similar-sized multilingual transformer models tasked with either masked language modeling - BERT (Multilingual) (Devlin et al.), RoBERTa XLM (Conneau et al.), Facebook XMOD (Pfeiffer et al.) - or decoder-only models - GPT2 (Radford et al.), Lllama 2 (Touvron et al.), and mGPT (Shliazhko et al.).

3 Results

For all models, accuracy on sentences without an interjection was 72% as opposed to 54% in sentences with an interjection, indicating a stark challenge in more complex cases. Additionally, there was a large drop in accuracy for the **GEN** test condition versus **NOM** (Figure 2), indicating that they likely assume the nominative case as the default form for nouns.

Moreover, in Figure 3 we see that, for all models, in the **GEN** condition, the accuracy is considerably larger with GEN as an attractor, and the lowest with the NOM and ACC attractors[3]. This shows that the models are using a "last occurrence" or case *agreement* strategy, by matching the most recent constituent, instead of finding the predicate of the negation.

In Figure 3, we see that Russian and Ukrainian generally performs better than Lithuanian and Latvian. The data for LLaMA 2, for example, correlates somewhat with the number of Wikipedia pages for each language[4].

[3]This pattern was inverted, though not as significant, for the **NOM** condition.

[4]Wikipedia entries were the only intentionally foreign language part of the LLaMA dataset

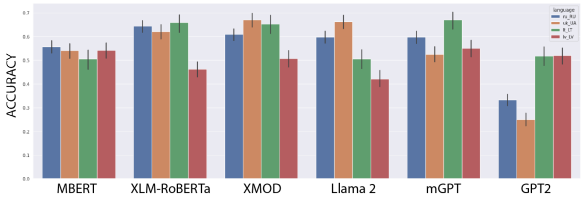

Figure 4: Accuracy by language (blue: Russian, orange: Ukrainian, green: Lithuanian, red: Latvian)

The models that did more similarly in all languages have provisions for low-resource language learning. XLM improves over mBERT with a reweighting of the Common Crawl dataset. mGPT, a fine-tuning of GPT3, is trained on a typologically weighted set of languages to improve overall performance. XMOD further improves over XLM by establishing language-specific modules in the later layers of the transformer, toggled by a language setting. In fact, a maximum Euclidean distance analysis on attention heads in XMOD between the two conditions identified two heads that were informative in case[5]. Ablating them drove down the accuracy to 51%. Crucially, all languages suffered form this ablation, that was identified only from Latvian, indicating that the model has learned similar cross-lingual syntactic patterns in the Balto-Slavic languages.

4 Discussion

This study highlights key insights and areas of improvement in transformer models' case processing in Balto-Slavic languages. There is a significant impact of training data on performance in similar-sized models, emphasizing the importance of balanced datasets. Additionally, the success of models like XMOD suggests that architectural innovations are very promising for enhancing performance in linguistically complex cases. All tested models exhibit a strategy balancing the "last occurrence" heuristic and long-range syntactic dependencies, suggesting an inherent trait of the transformer architecture. Additionally, attention analysis indicates an ability to learn shared syntactic rules across similar languages, showing the cross-lingual potential in enhancing multilingual model capabilities. These results highlight the need for continued evaluation on diverse languages.

[5]Head 6 in layer 4 and head 11 in layer 2

References

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Daniel Edmiston. 2020. A systematic analysis of morphological content in BERT models for multiple languages. *CoRR*, abs/2004.03032.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1195–1205, New Orleans, Louisiana. Association for Computational Linguistics.

Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. 2020. A systematic assessment of syntactic generalization in neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1725–1744, Online. Association for Computational Linguistics.

Anastasia Kobzeva, Suhas Arehalli, Tal Linzen, and Dave Kush. 2023. Neural networks can learn patterns of island-insensitivity in Norwegian. In *Proceedings of the Society for Computation in Linguistics 2023*, pages 175–185, Amherst, MA. Association for Computational Linguistics.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. *CoRR*, abs/1808.09031.

Kurt Micallef, Albert Gatt, Marc Tanti, Lonneke van der Plas, and Claudia Borg. 2022. Pre-training data quality and quantity for a low-resource language: New corpus and BERT models for Maltese. In *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing*, pages 90–101, Hybrid. Association for Computational Linguistics.

Roser Morante and Eduardo Blanco. 2021. Recent advances in processing negation. *Natural Language Engineering*, 27(2):121–130.

Gabriel Nicholas and Aliya Bhatia. 2023. Lost in translation: Large language models in non-english content analysis.

OpenAI. 2023. Gpt-4 technical report.

Jonas Pfeiffer, Naman Goyal, Xi Lin, Xian Li, James Cross, Sebastian Riedel, and Mikel Artetxe. 2022. Lifting the curse of multilinguality by pre-training modular transformers. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3479–3495, Seattle, United States. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Shauli Ravfogel, Yoav Goldberg, and Tal Linzen. 2019. Studying the inductive biases of RNNs with synthetic variations of natural languages. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3532–3542, Minneapolis, Minnesota. Association for Computational Linguistics.

Shauli Ravfogel, Francis M. Tyers, and Yoav Goldberg. 2018. Can LSTM learn to capture agreement? the case of basque. *CoRR*, abs/1809.04022.

Charlotte Rochereau, Benoît Sagot, and Emmanuel Dupoux. 2019. Modeling german verb argument structures: Lstms vs. humans. *CoRR*, abs/1912.00239.

Oleh Shliazhko, Alena Fenogenova, Maria Tikhonova, Vladislav Mikhailov, Anastasia Kozlova, and Tatiana Shavrina. 2022. mgpt: Few-shot learners go multilingual. *arXiv preprint arXiv:2204.07580*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Alex Warstadt, Yu Cao, Ioana Grosu, Wei Peng, Hagen Blix, Yining Nie, Anna Alsop, Shikha Bordia, Haokun Liu, Alicia Parrish, Sheng-Fu Wang, Jason Phang, Anhad Mohananey, Phu Htut, Paloma Jeretic, and Samuel Bowman. 2019. Investigating bert's knowledge of language: Five analysis methods with npis. pages 2870–2880.

Ethan Wilcox, Roger Levy, Takashi Morita, and Richard Futrell. 2018. What do RNN language models learn about filler–gap dependencies? In

Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pages 211–221, Brussels, Belgium. Association for Computational Linguistics.

Ethan Gotlieb Wilcox, Richard Futrell, and Roger Levy. 2023. Using Computational Models to Test Syntactic Learnability. *Linguistic Inquiry*, pages 1–44.

A Data Availability

As of May 10th, all the raw data is available here. The full jupyter notebook is available online here. The notebook was run with Google Colab Pro and can be run with 51GB of RAM.