Exploring Robustness of Multilingual LLMs on Real-World Noisy Data

Amirhossein Aliakbarzadeh[♣] and **Lucie Flek**[♣♠] and **Akbar Karimi**[♣♠]

[♣]Conversational AI and Social Analytics (CAISA) Lab, University of Bonn, Germany
[♠]Lamarr Institute for Machine Learning and Artificial Intelligence, Germany
amir.akbarzadeh95@gmail.com
{flek, ak}@bit.uni-bonn.de

Abstract

The textual content that is fed to Large Language Models (LLMs) can sometimes be noisy containing misspelled words. We investigate the effect of such mistakes on the performance of 9 language models, with parameters ranging from 0.2B to 13B, in 3 different NLP tasks, namely Natural Language Inference (NLI), Name Entity Recognition (NER), and Intent Classification (IC). We perform our experiments on 6 different languages (English, German, French, Spanish, Hindi, and Turkish) and build a dictionary of real-world noise for them using Wikipedia edit history. We show that the performance gap of the studied models on the clean and noisy test data averaged across all the datasets and languages ranges from **2.3** to **4.3** absolute percentage points. In addition, mT5 models, in general, show more robustness compared to BLOOM, Falcon, and BERT-like models. In particular, mT5 (13B), was the most robust on average overall, across the 3 tasks, and in 4 out of the 6 languages.

Introduction

Large Language Models (LLMs) have shown great performance in understanding a variety of languages. However, user interactions are not always error-free. Typographical errors can happen while the users interact with these models, leading to words that can not be found in their vocabulary. Furthermore, LLMs demonstrate vulnerability when exposed to textual noise in their input (Srivastava et al., 2020; Moradi and Samwald, 2021; Wang et al., 2023; Srivastava et al., 2020; Cai et al., 2022; Almagro et al., 2023; Stickland et al., 2023). This raises the critical question of how well these models perform in the presence of real-world noise. Additionally, given the multilingual nature of LLM interactions, understanding how performance varies across different languages is essential. Although large multilingual models perform impressively on various tasks and languages, their performance is usually degraded in non-English languages, especially low-resource ones (Etxaniz et al., 2023). Finally, the wide range of LLM sizes, from millions to billions of parameters, prompts the question of whether larger models are inherently more robust to real-world noise than smaller ones. We address these questions by first building a corpus of real-world noise from Wikipedia and then evaluating nine multilingual LLMs on the noisy data.

Method

We construct a collection of real-world typos from Wikipedia edits history (called WikiTypo) which contains around 40000 typos in 6 languages. To identify the typos and their corrections, first, we consider each article and its revisions and parse the main pages using the BeatutifulSoap library1. Then, we filter out the added and removed words on each page and extract a pair of words that 1) have one character-level Levenshtein edit distance from each other; 2) do not contain any number of special characters; and 3) have at least two characters each. The result is a dictionary of misspelled words with their corresponding correct spelling. This collection then is used to construct the noisy test sets for XNLI (Conneau et al., 2018) and SNIPS (Coucke et al., 2018) datasets. We use NLPAug (Ma, 2019) augmenter to create noisy data for the WikiANN (Rahimi et al., 2019) dataset since some sentences in this dataset contain only 2 or 3 words and most of the words are person, location, or organization names. We randomly replace words in a sentence with their incorrect version from the noise dictionary with two main parameters:

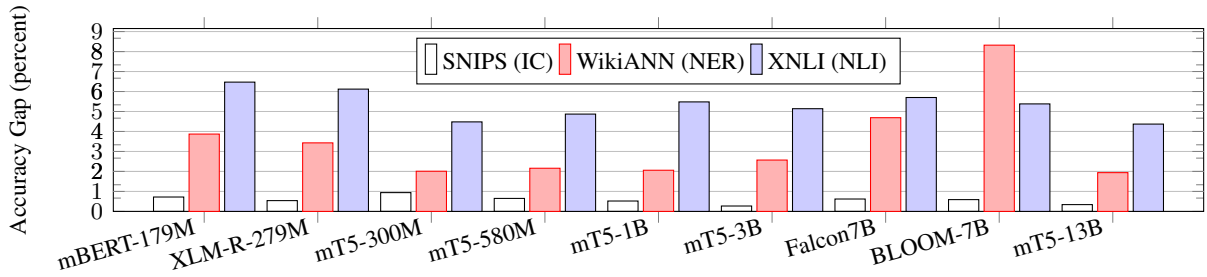

Figure 1: Average gap (in percentage points) between the accuracy of the experimented models on the clean data and the noisy data. The numbers indicate the average gap over all six languages on SNIPS, Wikiann, and XNLI datasets.

1) maximum number of augmented words ($m = 4$); 2) the proportion of the sentence ($r = 0.2$) to be changed.

We evaluate the robustness of multilingual LLMs against real-world noisy inputs by fine-tuning them on multilingual datasets and assessing their performance on clean and noisy test sets across three NLP tasks: Natural Language Inference (NLI), Named Entity Recognition (NER), and Intent Classification (IC). The studied models include encoder-only models (mBERT-179M Devlin et al. (2018), XLM-R-279M Conneau et al. (2019)), decoder-only models (BLOOM-7B Le Scao et al. (2022) , Falcon-7B Almazrouei et al. (2023)), and encoder-decoder models (mT5-(300M,580M,1B,3B,13B) Xue et al. (2020)).

Results

We fine-tune the models on WikiANN (NER) and SNIPS (IC) three times and on the XNLI (NLI) two times (due to the size of this dataset which is computationally expensive) and choose the best model based on its clean performance for evaluating on the test sets. We investigate the performance gap of the clean and noisy test sets for models and languages and the averaged values. Figure 1 compares the average performance gap over the languages for the individual datasets. We can see that the performance decrease is similar for all the models in the NLI and IC tasks, with the decrease being around 5 percentage points (±1) for the former and less than one percentage point for the latter. Additionally, BLOOM and Falcon models' performance gaps suggest that decoder-only models can have difficulty with some tasks such as named entity recognition compared to mBERT and XLM-R which use masked language modeling.

Looking at the mT5 models which have the same components with only difference in size, we see that the largest model (mT5-13B) is also the most robust. Our findings also indicate that English noisy test sets, particularly for XNLI, contain a higher frequency of noise, significantly impacting overall results. Unlike English, Turkish exhibits the least vulnerability to noise on XNLI and WikiANN datasets. Although the models perform poorly on English the gap converges to the average performance gap as the model size increases. Finally, our findings suggest that fine-tuning exclusively on noisy data yields a significant reduction in the performance gap at the cost of slightly decreasing the overall performance.

Conclusion

We build a corpus of real-world noise (typos) from Wikipedia edits history and explore how the noisy data impacts the performance of LLMs across languages. Performance gap across six languages indicates that all models exhibit vulnerability when encountering noisy input. In addition, our findings suggest that the robustness of language models against noisy input is influenced by various factors, including the size and language coverage of their training data, their underlying architectural design and parameter count, and the specific task they are evaluated on. While mT5 models demonstrate the best performance against noise likely due to their massive training data, size was not the sole factor. Architecture-wise, decoder-only models such as BLOOM and Falcon show more vulnerability to noise, especially in the NER task.

References

Mario Almagro, Emilio Almazán, Diego Ortego, and David Jiménez. Lea: Improving sentence similarity robustness to typos using lexical attention bias. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 36–46, 2023.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, et al. The falcon series of open language models. *arXiv preprint arXiv:2311.16867*, 2023.

Shanqing Cai, Subhashini Venugopalan, Katrin Tomanek, Ajit Narayanan, Meredith Morris, and Michael Brenner. Context-aware abbreviation expansion using large language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1261–1275, 2022.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, 2018.

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

Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, et al. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. *arXiv preprint arXiv:1805.10190*, 2018.

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

Julen Etxaniz, Gorka Azkune, Aitor Soroa, Oier Lopez de Lacalle, and Mikel Artetxe. Do multilingual language models think better in english? *arXiv preprint arXiv:2308.01223*, 2023.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. Bloom: A 176b-parameter open-access multilingual language model. 2022.

Edward Ma. Nlp augmentation. https://github.com/makcedward/nlpaug, 2019.

Milad Moradi and Matthias Samwald. Evaluating the robustness of neural language models to input perturbations. *arXiv preprint arXiv:2108.12237*, 2021.

Afshin Rahimi, Yuan Li, and Trevor Cohn. Massively multilingual transfer for ner. *arXiv preprint arXiv:1902.00193*, 2019.

Ankit Srivastava, Piyush Makhija, and Anuj Gupta. Noisy text data: Achilles' heel of bert. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 16–21, 2020.

Asa Cooper Stickland, Sailik Sengupta, Jason Krone, He He, and Saab Mansour. Robustification of multilingual language models to real-world noise in crosslingual zero-shot settings with robust contrastive pretraining. In *17th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2023*, pages 1367–1383. Association for Computational Linguistics (ACL), 2023.

Haoyu Wang, Guozheng Ma, Cong Yu, Ning Gui, Linrui Zhang, Zhiqi Huang, Suwei Ma, Yongzhe Chang, Sen Zhang, Li Shen, et al. Are large language models really robust to word-level perturbations? *arXiv preprint arXiv:2309.11166*, 2023.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*, 2020.