УДК 004.81

ИССЛЕДОВАНИЕ ОСОБЕННОСТЕЙ ИСПОЛЬЗОВАНИЯ LLM В ЗАДАЧЕ CODE-SWITCHING В ЗАДАНИЯХ ДЛЯ ИЗУЧЕНИЯ ИНОСТРАННЫХ ЯЗЫКОВ

Мазеин Н.О. (Университет ИТМО, студент бакалавриата),
Тихонова К.Е. (Университет ИТМО, студентка бакалавриата),
Насыров Н.Ф. (Университет ИТМО, студент магистратуры),
Федоров Д.А. (Университет ИТМО, кандидат технический наук, доцент)

Введение

Code-switching (переключение/смешение кодов) – распространенное явление среди многих многоязычных сообществ, когда говорящие переходят от одного языка к другому в рамках одного высказывания. В контексте исследования особенностей использования больших языковых моделей (LLM – Large Language Model) в задаче code-switching в заданиях для изучения иностранных языков, становится ясно, что LLM может быть ключевым инструментом для искусственного создания материалов с применением code-switching.

Использование LLM в исследовании задач code-switching позволяет более глубоко и точно моделировать переходы между языками, выявлять паттерны и зависимости, учитывать сходства и различия в грамматике языков. Это имеет важное значение для процесса обучения людей иностранным языкам, поскольку code-switching является частью языкового опыта человека, и эффективное его моделирование может улучшить качество обучения [1] и повысить межкультурное понимание [2]. Стоит упомянуть, что альтернативой чтения текстов со смешением кодов в рамках изучения иностранных языков является чтение текстов, написанных по методу Ильи Франка, которые показали свою эффективность в расширении словарного запаса и облегчить процесс адаптации к чтению иностранной литературы [3].

В данной работе будет рассмотрено несколько моделей LLM в рамках исследования возможностей и ограничений этих моделей для использования в генерации текстов с переключением кодов в целях последующего применения в изучении иностранных языков.

Основная часть

Изучение code-switching – важная задача NLP исследователей, одной из основных проблем которой является отсутствие качественных датасетов с разметкой типа code-switching [4], особенно для языков с малой зоной распространения носителей. Поэтому с появлением мультиязычных больших языковых моделей одним из направлений исследований стала генерация текстов с переключением кодов при помощи LLM. Однако, в отличие от людей, не все мультиязычные LLM способны применять code-switching ввиду архитектурных особенностей [5] и малой представленности подобных текстов в тренировочных датасетах. Интерес представляет и то, что некоторые мультиязычные модели после fine-tune уступают ChatGPT в качестве сгенерированных текстов со смешанными кодами.

В рамках данной работы мы провели исследование возможностей 20 открытых предобученных больших языковых моделей из списка представленных на данный момент актуальных LLM [6]. Примеры промптов использованных для тестирования:

1) Write a code-mixed text in English and Russian about traveling around the world, use complex English and Russian grammar, code-switch at least 5 times per sentence,
2) [Английский текст сложности C2] Rewrite this text to English-Russian code-switching text, use at least 3 code-switches per sentence,
3) Rewrite this text but switch words ([], [], [], … – английские слова или словосочетания) to their Russian counterparts.

В то время, как большинство моделей не справились ни с одной из поставленных задач, было получено несколько интересных результатов, представленных ниже.

1) Самой способной моделью из протестированных оказалась gpt-4-1106-preview. Она с успехом справлялась со всеми поставленными задачами, заменяла слова абсолютно

осмысленно и согласно контексту (в отличие от того же mistral-medium, который переводил слова, не обращая внимания на контекст).

2) ChatGPT (gpt-3.5-turbo) приблизился по качеству генерации к своей более продвинутой версии, справляясь с поставленными задачами из промптов, уступая лишь в том, что был неспособен осуществлять частый code-switching, зачастую переставая делать замену для целых предложений или даже абзацев, в то время как gpt-4-1106-preview справлялся даже когда его просили менять код через каждое слово.

3) mixtral-8x7b-instruct-v0.1 оказался самым продвинутым в линейке mistral, однако был не способен осуществлять code-switching, вместо этого он писал текст по методу Ильи Франка, записывая перевод в скобках после каждого предложения или сложных словосочетаний.

4) llama-2-70b-chat оказалась единственной моделью, которая реализовала задачу code-switching заменяя английские слова на русский транслит, то есть слово hello превращалось в privet. Эта странная особенность может быть списана на хорошую адаптацию моделей LLaMA в изучении языков и глубокое понимание их структуры. Например, LLaMA модели легко подстраиваются под новые языки при fine-tuning'е и могут быть использованы для code-switching'а между несколькими языками с различной структурой без потери смысла внутри контекста [7].

Заключение

В настоящий момент задача code-switching'а представляет особый интерес с точки зрения обучения иностранным языкам, поскольку является недостаточно изученной, при этом потенциально ее имплементация в процесс обучения может привести к усвоению вокабуляра с повышенной эффективностью, сократив при этом культурную разницу и улучшив понимание между носителями языка и изучающими язык, поскольку является не просто задачей точечного перевода слов внутри предложения, но задачей анализа структуры языка и мышления носителя. С точки зрения NLP данная задача урезана фактом дефицита качественных датасетов с разметкой code-switching. В данной проблеме исследователям потенциально могут быть полезны мультиязычные большие языковые модели, которые учатся анализировать структуры разных языков и выстраивать предложения на основе освоенных структур. Не исключено, что в будущем данная задача будет качественно решена при помощи LLM, специально обученной находить связи между структурами разных языков.

Список используемых источников

1. Tashmatova Gulnara Rafailovna Code-switching in teaching a foreign language // Проблемы Науки, 2020, №3 (148), URL: https://cyberleninka.ru/article/n/code-switching-in-teaching-a-foreign-language
2. Seymen-Bilgin, Sezen Code switching in English language teaching (ELT) teaching practice in Turkey: Student teacher practices, beliefs and identity // Educational Research and Reviews - Kocaeli, Turkey : AcademicJournals, Vol. 11, 2016. - PP. 686-702, DOI: 10.5897/ERR2016.2802.
3. Buranova, Madina Linguistic analysis of reading and teaching reading by literary translation // Общество и инновации, Vol. 1, 2020. - PP. 501-505, DOI: 10.47689/2181-1415-vol1-iss1/s-pp501-505.
4. Yong, Z.X., Zhang, R., Forde, J.Z., Wang, S., Cahyawijaya, S., Lovenia, H., Sutawika, L., Cruz, J.C.B., Phan, L., Tan, Y.L. and Aji, A.F. Prompting Large Language Models to Generate Code-Mixed Texts: The Case of South East Asian Languages // arXiv preprint, arXiv:2303.13592, 2023
5. Ruochen Zhang, Samuel Cahyawijaya, Jan Christian Blaise Cruz, Genta Winata, Alham Aji Multilingual Large Language Models Are Not (Yet) Code-Switchers // Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. - Singapore: Association for Computational Linguistics, 2023. - PP. 12567–12582, DOI: 10.18653/v1/2023.emnlp-main.774
6. Chatbot Arena // URL: https://chat.lmsys.org/

7. Zhao, J., Zhang, Z., Zhang, Q., Gui, T. and Huang, X. arXiv preprint arXiv:2401.01055. Llama beyond english: An empirical study on language capability transfer // arXiv:2401.01055. - 2024, DOI: https://doi.org/10.48550/arXiv.2401.01055

Мазеин Н.О. (автор)	Подпись
Тихонова К. Е. (автор)	Подпись
Насыров Н. Ф. (консультант)	Подпись
Федоров Д. А. (научный руководитель)	Подпись