Сравнительный анализ качества автоматической морфоразметки для языков со значительной разницей в индексе агглютинации (на примере русского и турецкого)

Анастасия А. Российская
Российский государственный гуманитарный университет,
Москва, Россия, flxgtm@gmail.com

Данная работа посвящена анализу качества автоматической морфоразметки для языков со значительной разницей в индексе агглютинации (в терминах квантитативной типологии, см. [Greenberg 1960, pp. 178–194; Касевич, Яхонтов 1982]) на примере русского и турецкого языков. Было проведено исследование на материале, самостоятельно выкачанном из социальной сети Twitter. Основной гипотезой нашего исследования было то, что правиловые парсеры достаточно хорошо работают для языков с высоким индексом агглютинации, или по крайней мере лучше, чем для языков с низким. Высокая степень морфологической регулярности давала нам основания ожидать в том числе, что правиловые парсеры покажут на турецком материале качество, сравнимое с нейронными парсерами для языков со схожей русскому морфологической системой.

Морфоразметка, или морфологический парсинг — это процесс анализа морфемного состава словоформы. Соответственно, морфологический парсер — это программа, осуществляющая этот процесс автоматически. Морфологические парсеры можно разделить на правиловые и нейронные.

Правиловые парсеры — это парсеры, в основе работы которых лежат правила, прописанные разработчиком-лингвистом; парсеры на нейронных сетях используют обученные на размеченных данных модели, которые самостоятельно без помощи человека пытаются установить правила. Предположительно, человек не в состоянии предусмотреть все случаи, и поэтому обычно парсеры на нейронных сетях показывают гораздо более высокие результаты.

Для морфологической системы русского языка характерно большое количество отклонений от изоморфизма, таких как фузия, кумуляция, супплетивизм и синкретизм, соответственно, можно предположить, что эта система сложнее поддается описанию с помощью правил, нежели морфология языков с высоким индексом агглютинации, таких, как турецкий.

При анализе материала были использованы парсер MyStem[1] (снятие омонимии которого работает на классическом алгоритме машинного обучения) и правиловый парсер PyMorphy[2] для русского языка, правиловые парсеры Zeyrek[3] и FsmMorph[4] для турецкого языка и парсер UDPipe[5] на нейронных сетях для обоих языков. Для русского языка использовалась модель ru_core_news_sm[6], обученная на новостных текстах. Для турецкого языка использовалась модель UDPipe, обученная на датасете IMST[7], который также содержит новостные и художественные тексты.

В качестве материала для тестирования выбранных парсеров были использованы твиты двух открытых аккаунтов: одного на русском, другого на турецком. Выбор текстов соцсети, отличающихся по стилю от датасета, на котором обучались выбранные нейронные парсеры, был намеренным — мы ставили целью проверить качество морфоразметки на сложных текстах.

Общее количество подвергнутого анализу материала — около 7200 токенов.

Статистика по токенам в русском и турецком языках представлена в таблицах ниже.

	UDPipe	FsmMorph	Zeyrek
верно разобранные токены	72%	63%	62%
нераспознанные токены	17%	21%	26%
ошибки	11%	16%	12%

Таблица 1. Статистика результатов морфоразметки по токенам для турецкого

[1] https://github.com/nlpub/pymystem3
[2] https://github.com/kmike/pymorphy2
[3] https://github.com/obulat/zeyrek
[4] https://github.com/starlangsoftware/TurkishMorphologicalAnalysis-Py
[5] https://github.com/ufal/udpipe
[6] https://spacy.io/models/ru#ru_core_news_sm. Выбор пал на данную модель, поскольку ожидалось, что даже нейронная сетка, обученная на маленьком датасете, состоящем из текстов другого типа, будет показывать лучшее качество, чем правиловый парсер.
[7] https://universaldependencies.org/treebanks/tr_imst/index.html

	UDPipe	PyMorphy	MyStem
верно разобранные токены	73%	83%	80%
нераспознанные токены	7%	6%	6%
ошибки	20%	11%	14%

Таблица 2. Статистика результатов морфоразметки по токенам для русского

Как видно по таблице для турецкого языка, парсер на нейронных сетях продемонстрировал лучший результат — 72% верно разобранных токенов. Однако у всех парсеров, использованных для турецкого, процент нераспознанных токенов значительно выше по сравнению с русским. Это связано с тем, что пользователи, ведущие соцсети на турецком, часто опускают диакритики, используя латиницу вместо букв современного турецкого алфавита.

Тот факт, что UDPipe показал самый маленький процент верно распознанных токенов для русского объясняется тем, что этот парсер был обучен на датасете с новостными текстами, как упоминалось ранее, а в нашем исследовании был использован для анализа текстов соцсетей.

Таким образом, главное преимущество нейронных парсеров для языков с высоким индексом агглютинации типа турецкого заключается в том, что они справляются с отсутствием диакритик и опечатками. Вопрос о связи относительного качества работы правиловых парсеров с индексом агглютинации размечаемого языка остаётся открытым, потому как особенности графики Twitter не были приняты во внимание и значительно повлияли на результаты.

UDPipe и для русского, и для турецкого даёт примерно похожее качество, и с диакритиками справляется, следовательно, можно предположить, что, если бы пользователями использовались верные буквы, процент верно разобранных токенов на турецком мог быть ещё выше.

На основании вышеизложенного можно заключить, что по крайней мере для разметки текстов соцсетей правиловый парсер недостаточен даже при высоком индексе агглютинации анализируемого языка. Это связано с тем, что правиловые парсеры не умеют учитывать ошибки (опечатки, потерянные диакритики,

окказионализмы), в то время как парсеры на нейронных сетях справляются с этой проблемой [Ahmed et al. 2022, pp. 6–7].

Литература

Касевич, Яхонтов 1982 – Квантитативная типология языков Азии и Африки. Отв. ред. В. Б. Касевич, С. Е. Яхонтов. Ленинградский университет. 1982.

Ahmed et al. 2022 – S. Ahmed et al. Tafsir Dataset: A Novel Multi-Task Benchmark for Named Entity Recognition and Topic Modeling in Classical Arabic Literature. 6-7. 2022.

Greenberg 1960 – J. H. Greenberg. A Quantitative Approach to the Morphological Typology of Language: International Journal of American Linguistics 26. 178-194. 1960.