DOI: 10.15514/ISPRAS-2022-34(6)-14

Оценка языковой способности нейронных моделей на материале предикативного согласования в русском языке

К.А. Студеникина, ORCID: 0000-0002-4098-7167 <xeanst@gmail.com>
Московский государственный университет имени М.В. Ломоносова,
119991, Россия, Москва, Ленинские горы, д. 1

Аннотация. Исследование нацелено на то, чтобы изучить способность нейронных сетей моделировать грамматику, которая проявляется в функции автоматической оценки грамматичности языковых выражений. В данной работе моделируются правила предикативного согласования по числу в русском языке. Для обучения языковых моделей был создан датасет, включающий искусственно сгенерированные грамматичные и неграмматичные предложения. Мы используем трансферное обучение предобученных нейронных сетей. Результаты показывают, что все рассмотренные модели демонстрируют высокие результаты при дообучении на задачу оценки грамматичности. Точность классификации снижается для предложений с неодушевленными существительными, поскольку для них, в отличие от одушевленных существительных, наблюдается совпадение форм именительного и винительного падежа. Сложность синтаксической структуры оказывается значимой для русскоязычных моделей и модели для славянских языков, но не влияет на распределение ошибок для мультиязычных моделей.

Ключевые слова: языковая способность; языковые модели; трансферное обучение; искусственный интеллект; обработка естественного языка; оценка грамматичности

Для цитирования: Студеникина К.А. Оценка языковой способности нейронных моделей на материале предикативного согласования в русском языке. Труды ИСП РАН, том 34, вып. 6, 2022 г., стр. 179-184. DOI: 10.15514/ISPRAS-2022-34(6)-14

Благодарности. Исследование выполнено при финансовой поддержке Некоммерческого Фонда развития науки и образования «Интеллект».

Evaluation of neural models' linguistic competence: evidence from Russian predicate agreement

K.A. Studenikina, ORCID: 0000-0002-4098-7167 <xeanst@gmail.com>
Lomonosov Moscow State University,
GSP-1, Leninskie Gory, Moscow, 119991, Russia

Abstract. This study investigates the linguistic competence of modern language models. Artificial neural networks demonstrate high quality in many natural language processing tasks. However, their implicit grammar knowledge remains unstudied. The ability to judge a sentence as grammatical or ungrammatical is regarded as key property of human's linguistic competence. We suppose that language models' grammar knowledge also occurs in their ability to judge the grammaticality of a sentence. In order to test neural networks' linguistic competence, we probe their acquisition of number predicate agreement in Russian. A dataset consisted of artificially generated grammatical and ungrammatical sentences was created to train the language models. Automatic sentence generation allows us to test the acquisition of particular language phenomenon, to detach from vocabulary and pragmatic differences. We use transfer learning of pre-trained neural networks. The results show that all the considered models demonstrate high accuracy and Matthew's correlation coefficient values

which can be attributed to successful acquisition of predicate agreement rules. The classification quality is reduced for sentences with inanimate nouns which show nominative-accusative case syncretism. The complexity of the syntactic structure turns out to be significant for Russian models and a model for Slavic languages, but it does not affect the errors distribution of multilingual models.

Keywords: linguistic competence; language models; transfer learning; artificial intelligence; natural language processing; grammaticality judgments.

For citation: Studenikina K.A. Evaluation of neural models' linguistic competence: evidence from Russian predicate agreement. Trudy ISP RAN/Proc. ISP RAS, vol. 34, issue 6, 2022. pp. 179-184 (in Russian). DOI: 10.15514/ISPRAS-2022-34(6)-14

Acknowledgements. This work was supported by Non-commercial Foundation for support of Science and Education «INTELLECT».

1. Введение

Современные нейронные сети демонстрируют уровень обработки естественного языка, сравнимый с человеческим. Появляется все больше новых языковых моделей различной архитектуры, в связи с чем возникает необходимость их сравнения. В науке о данных уделяется большое внимание созданию так называемых бенчмарков – ресурсов для обучения, оценки и анализа языковых моделей. Подобный бенчмарк существует и для русского языка под названием Russian SuperGLUE [1]. Он включает в себя преимущественно такие задачи, которые оценивают понимание смысла текста: это, например, разрешение семантической неоднозначности [2], логический вывод по тексту [3], поиск ответа на вопрос в тексте [4].

Тем не менее, остается открытым вопрос о том, обладают ли нейронные сети имплицитным знанием грамматики. Умение оценивать грамматическую правильность предложения на родном языке является ключевым свойством языковой способности человека [5]. Суждения о правильности языкового выражения получили название «оценок грамматичности/приемлемости». В лингвистике суждения о грамматичности, порожденные людьми, выступают как эмпирическая база для моделирования языковой способности человека. Следовательно, возможно использовать автоматические суждения о грамматичности для оценки языковой способности нейронных моделей.

Предшествующие исследования в данном направлении выполнены преимущественно на материале английского языка, для русского языка работ значительно меньше. Первое подобное исследование решает задачу автоматической оценки приемлемости в русском языке на основе корпуса предложений из лингвистических статей [6], что вызывает две проблемы. Во-первых, данные не содержат информацию, какой тип ошибки наблюдается в предложении, что делает невозможной детальную оценку языковой способности модели. Во-вторых, используется бинарная классификация, однако примеры иллюстрируют сложные феномены, которые не могут быть однозначно оценены как грамматичные или неграмматичные.

Задача нашего исследования состоит в том, чтобы изучить способность нейронных сетей к моделированию грамматики на материале конкретного феномена: функции автоматической оценки грамматичности языковых выражений на материале русского языка. Для исследования будут использованы искусственно сгенерированные данные, не содержащие пограничных случаев, что делает возможной бинарную классификацию по грамматичности. Проведенные исследования и код с результатами обучения представлены по ссылке: https://github.com/Xeanst/Grammaticality-judgements-with-neural-language-models.

2. Автоматическая генерация данных

Цель исследования состоит в том, чтобы оценить, насколько языковые модели обладают знаниями грамматики. Поскольку нейросеть должна разграничить граммматичные и неграмматичные предложения, обучающие и тестовые данные содержат не только корректные примеры, но и примеры с ошибками. В качестве датасета могут быть использованы искусственно сгенерированные предложения [7, 8]. Данный метод позволяет самостоятельно определять, ошибка на какое языковое явление будет содержаться в предложении. Однако сами примеры будут несколько неестественными. В нашем исследовании мы осуществили искусственную генерацию предложений. Поскольку нам необходимо понять, насколько хорошо модель «понимает» устройство грамматики языка, некоторая неестественность предложений поможет абстрагироваться от лексических и прагматических аспектов и оценить именно знание грамматики.

Было сгенерировано 700 примеров. Мы исследовали, насколько хорошо модель способна усвоить механизм согласования по числу подлежащего и сказуемого: в грамматичных примерах число подлежащего и сказуемого совпадает (1), в неграмматичных – отличается (2). Для каждого предложения сгенерировано несколько вариантов, содержащих одни и те же лексические единицы и отличающиеся ровно двумя параметрами, числом подлежащего и числом сказуемого. Это позволяет определить, насколько хорошо языковая модель усваивает целевое грамматическое явление – предикативное согласование.

(1) a. Студент_ рассуждал_.
  b. Студенты рассуждали.
(2) a. Студент_ рассуждали.
  b. Студенты рассуждал_.

Мы рассматривали два лингвистических аспекта, которые могут осложнить автоматическое моделирование правил предикативного согласования. В первую очередь, это параметр одушевленности: одна половина примеров содержала одушевленные существительные, для которых формы именительного и винительного падежа различаются (3), другая – неодушевленные, где падежные формы совпадают (4). В первом случае модель должна была показать высокое качество классификации, однозначно определить подлежащее и дополнение предложения и правильно предсказать, в каких предложениях предикативное согласование корректно, а в каких – ошибочно. Во втором случае мы ожидали более низкое качество предсказания, поскольку по падежной форме дополнения нельзя определить, в каком падеже оно стоит – в именительном или винительном.

(3) a. Студент_ знал_ охранников.
  b. Студент_ знали охранников.
  c. Охранников знал_ студент_.
  d. Охранников знали студент_.
(4) a. Студент_ знал_ рассказы.
  b. Студент_ знали рассказы.
  c. Рассказы знал_ студент_.
  d. Рассказы знали студент_.

Помимо деления по одушевленности, данные содержали примеры различной степени сложности. Если в примере есть несколько существительных или несколько глаголов, то модели будет сложнее понять, грамматично ли предложение. Если же в предложении имеется только одно существительное и только один глагол, модель должна легко понимать, правильно ли указано число глагола. Всего данные содержали 7 типов синтаксических структур (по 10 предложений на каждый тип): простое предложение с подлежащим без дополнения (1-2), простое предложение с подлежащим и дополнением в прямом порядке (3a-b, 4a-b), простое предложение с подлежащим и дополнением в обратном порядке (3c-d, 4c-d), сложноподчиненное предложение (5), простое предложение с зависимым существительным при подлежащем (6), сложное предложение с субъектным зависимым предложением при подлежащем (7), сложное предложение с объектным зависимым предложением при подлежащем (8).

(5) a. Писатели говорили, что студенты рассуждали.

  b. Писатели говорили, что студенты рассуждал_.
(6) a. Студент_ рядом с писателями рассуждал_.
  b. Студент_ рядом с писателями рассуждали.
(7) a. Студент_, который знал_ охранников, рассуждал_.
  b. Студент_, который знал_ охранников, рассуждали.
(8) a. Студент_, которого знали охранники, рассуждал_.
  b. Студент_, которого знали охранники, рассуждали.

Таким образом, благодаря искусственной генерации примеров, удалось получить объемный сбалансированный датасет.

3. Обучение и тестирование моделей

В обучающей и тестовой выборке грамматичные предложения имели метку «1», предложения с ошибкой – метку «0». Следовательно, задача оценки грамматичности сводится к задаче бинарной классификации. В исследовании мы использовали трансферное обучение. Заранее обученные языковые модели с помощью полученного датасета дообучались на задачу классификации по грамматичности. Данные были разделены в следующем соотношении: для обучения – 80%, для валидации – 10%, для тестирования – 10%. При обучении количество эпох было равно 5, использовался графический процессор. Мы использовали как модели для русского языка: ruBERT [9] и Russian RoBERTa [10], так и мультиязычные модели: multilingual BERT (104 языка) [11], Slavic BERT (4 языка) [12], XLM RoBERTa (15 языков) [13].

В табл. 1 представлены общие результаты тестирования моделей в задаче классификации по грамматичности. Для оценки эффективности работы классификационного алгоритма используются такие метрики, как точность (accuracy) и коэффициент корреляции Мэтьюса (Matthews Correlation Coefficient, MCC). Можно заметить, что модели показывают довольно высокое качество. Наиболее высокие результаты демонстрируют модели на основе архитектуры BERT: модель для русского языка ruBERT, мультиязычная модель multilingual BERT и модель для славянских языков SlavicBERT. Более низкие результаты достигаются на основе архитектуры Roberta: модель для русского языка Russian RoBERTa и мультиязычная модель XLM Roberta.

Табл. 1. Общие результаты тестирования моделей
Table 1. General results for models testing

№	Модель	Точность	Коэффициент корреляции Мэтьюса
1	ruBERT multilingual BERT	98.6	0.97
2	Slavic BERT	97.1	0.94
3	Russian RoBERTa	77.1	0.54
4	XLM Roberta	57.1	0.33

Проведем лингвистический анализ результатов и подробнее рассмотрим точность классификации в зависимости от одушевленности существительных и синтаксической сложности. В табл. 2 представлены значения точности классификации для предложений с одушевленными и неодушевленными существительными. Результаты показывают, что все языковые модели оказались восприимчивы к совпадению падежных форм и продемонстрировали больше ошибок при определении корректной согласовательной стратегии для неодушевленных существительных, чем для одушевленных. Модели ruBERT, multilingual BERT и Slavic BERT ошибочно предсказали оценку только для предложений с неодушевленными существительными. Модели Russian RoBERTa и XLM Roberta также допустили большее количество ошибок на предложениях с неодушевленными существительными.

Табл. 2. Точность классификации при тестировании моделей на предложениях с одушевленными и неодушевленными существительными
Table 2. Classification accuracy of models testing on sentences with animate and inanimate nouns

Существительное	ruBERT	multilingual BERT	Slavic BERT	Russian RoBERTa	XLM Roberta
Одушевленное	100.0	100.0	100.0	80.0	62.9
Неодушевленное	97.1	97.1	94.3	74.3	51.4

В табл. 3 продемонстрирована точность классификации для предложений различной синтаксической сложности. Для некоторых моделей, действительно наблюдается больше ошибок для предложений со сложной синтаксической структурой: для сложноподчиненного предложения (тип 4, модель ruBERT), для простых предложений с обратным порядком подлежащего и дополнения (тип 3, модель Slavic BERT и модель Russian RoBERTa), для сложных предложений с объектным зависимым при подлежащем (тип 7, также модель Russian RoBERTa). В то же время, другие модели допускают ошибки вне зависимости от синтаксической сложности структуры: они демонстрируют наибольшее количество ошибок для простых предложений с подлежащим (тип 1, модели multilingual BERT и XLM Roberta).

Табл. 3. Точность классификации при тестировании моделей на предложениях с различной синтаксической структурой
Table 3. Classification accuracy of models testing on sentences with various syntactic structure

№	Тип предложения	ruBERT	multilingual BERT	Slavic BERT	Russian RoBERTa	XLM Roberta
1	Простое предложение с подлежащим без дополнения	100.0	90.0	100.0	90.0	30.0
2	Простое предложение с подлежащим и дополнением в прямом порядке	100.0	100.0	100.0	80.0	60.0
3	Простое предложение с подлежащим и дополнением в обратном порядке	100.0	100.0	80.0	60.0	50.0
4	Сложноподчиненное предложение	90.0	100.0	100.0	80.0	40.0
5	Простое предложение с зависимым существительным при подлежащем	100.0	100.0	100.0	70.0	70.0
6	Сложное предложение с субъектным зависимым предложением при подлежащем	100.0	100.0	100.0	100.0	80.0
7	Сложное предложение с объектным зависимым предложением при подлежащем	100.0	100.0	100.0	60.0	80.0

Наши ожидания относительно лингвистических аспектов, которые могут осложнять классификацию, подтвердились лишь частично. Для большинства моделей возникают трудности при оценке предложений с неодушевленными существительными, поскольку для них, в отличие от одушевленных существительных, наблюдается совпадение форм именительного и винительного падежа. Однако предположение относительно сложности структур подтвердилось не до конца. Синтаксическая структура оказывается значимой для

русскоязычных моделей и модели для славянских языков, но не влияет на распределение ошибок для мультиязычных моделей.

4. Заключение

В данной работе была проведена оценка языковой способности нейронных моделей на материале русского языка. Исследование показало, что при обучении на большом количестве неразмеченных данных модели некоторым образом выучивают грамматику. Полученный список моделей, ранжированный по точности классификации, имеет высокую технологическую ценность и может быть использован для выбора модели при решении различных задач обработки языка. Если модель хорошо справляется с автоматической оценкой грамматичности, она отлично покажет себя в таких задачах, как морфологический и синтаксический анализ.

Список литературы / References

[1] Shavrina T., Fenogenova A. et al. RussianSuperGLUE: A Russian language understanding evaluation benchmark. In Proc. of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 4717-4726.

[2] Panchenko A., Lopukhina A. et al. RUSSE'2018: a shared task on word sense induction for the Russian language. In Proc. of the International Conference “Dialogue 2018”, 2018, pp. 547-564.

[3] De Marneffe M. C., Simons M., Tonhauser J. The CommitmentBank: Investigating projection in naturally occurring discourse. Proceedings of Sinn und Bedeutung, vol. 23, issue 2, 2019, pp. 107-124.

[4] Glushkova T., Machnev A. et al. DaNetQA: A Yes/No Question Answering Dataset for the Russian Language. Lecture Notes in Computer Science, vol. 12602, 2020, pp. 57–68.

[5] Chomsky N. Aspects of the theory of syntax. MIT Press, 1969, 251 p.

[6] Mikhailov V., Shamardina T. et al. RuCoLA: Russian Corpus of Linguistic Acceptability. In Proc. of the 2022 Conference on Empirical Methods in Natural Language Processing, 2022, pp. 5207-5227.

[7] Marvin R., Linzen T. Targeted syntactic evaluation of language models. In Proc. of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018, pp. 1192-1202.

[8] Warstadt A., Parrish A. et al. BLiMP: The benchmark of linguistic minimal pairs for English. Transactions of the Association for Computational Linguistics, vol. 8, 2020, pp. 377-392.

[9] Kuratov Y., Arkhipov M. Adaptation of deep bidirectional multilingual transformers for Russian language. In Proc. of the International Conference “Dialogue 2019”, 2019, pp. 333-339.

[10] Blinov P., Avetisian M. Transformer models for drug adverse effects detection from tweets. In Proc. of the Fifth Social Media Mining for Health Applications Workshop & Shared Task, 2020, pp. 110-112.

[11] Devlin J., Chang M.W. et al. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proc. of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language, 2018, pp. 4171-4186.

[12] Arkhipov M., Trofimova M. et al. Tuning multilingual transformers for language-specific named entity recognition. In Proc. of the 7th Workshop on Balto-Slavic Natural Language Processing, 2019, pp. 89-93.

[13] Lample G., Conneau A. Cross-lingual language model pretraining. In Proc. of the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), 2019, 11 p.

Информация об авторах / Information about authors

Ксения Андреевна СТУДЕНИКИНА является программистом лаборатории автоматизированных лексикографических систем НИВЦ МГУ и аспирантом кафедры теоретической и прикладной лингвистики филологического факультета МГУ. Её научные интересы включают автоматическую обработку естественного языка и исследование синтаксиса русского языка.

Kseniia Andreevna STUDENIKINA is a programmer in the Laboratory for Computational Lexicography of Research Computing Center of MSU and PhD student at the Department of Theoretical and Applied Linguistics of Philological Faculty of MSU. Her research interests include natural language processing and studies of Russian syntax.