Автоматическая оценка грамматичности с помощью нейронных языковых моделей[1]

Студеникина Ксения, 2 курс магистратуры, филологический факультет МГУ

1. Введение

В лингвистике принято считать, что основным свойством языковой способности человека является возможность определять, насколько грамматически корректно предложение [Chomsky 1965]. Под грамматичностью в данном случае подразумевается не методическая правильность в соответствии со школьными знаниями, а именно интуитивная оценка, основанная на врожденной грамматике человека. Подобные суждения говорящих о правильности языкового высказывания получили название «оценок грамматичности». Лингвисты используют суждения о грамматичности для исследования синтаксической структуры предложений.

Поскольку человеческие суждения о грамматичности выступают как один из основных типов данных для моделирования языковой способности людей, возможно использовать автоматические суждения о грамматичности для оценки языковой способности нейронных моделей. Так, современные нейросетевые модели достигают высокой степени компетентности во многих прикладных задачах понимания естественного языка: анализ тональности, логический вывод по тексту, поиск ответа на вопрос в тексте. Цель данного исследования состоит в том, чтобы оценить, насколько языковые модели обладают знаниями грамматики. В работе используется трансферное обучение предобученных языковых моделей и дообучение их на задачу оценки грамматичности.

2. Литературный обзор: грамматичность vs. приемлемость

Предшествующие исследования, в которых решалась задача оценки языковой способности нейронных моделей, ставили перед собой цель автоматического предсказания оценок приемлемости, а не оценок грамматичности. Различия между грамматичностью и приемлемостью состоят в следующем. Под грамматичностью подразумевается грамматическая корректность предложения и отсутствие синтаксических ошибок, как выбор падежа, согласование по роду, лицу и числу. Грамматичность является бинарной категорией: предложение может быть либо грамматичным, либо неграмматичным. Приемлемость же является более широким понятием. Предложение считается приемлемым, если оно корректно не только грамматически, но и семантически (осмысленно и логично), и стилистически (отсутствуют речевые ошибки). Так, известное предложение (1a) (в оригинале (1b), [Chomsky 1957]) является грамматичным, но неприемлемым в силу отсутствия какого-либо смысла, то есть семантической некорректности. Аналогично, предложение (2) будет грамматичным, но неприемлемым, поскольку слово *ихний* на данном этапе развития языка считается просторечным [Евгеньева 1999]. Тем не менее, язык развивается, и в будущем это слово может потерять разговорно-сниженный оттенок – тогда предложение будет и грамматичным, и приемлемым.

[1] Данные исследования и код с результатами обучения представлены по ссылке: https://github.com/Xeanst/Grammaticality-judgements-with-neural-language-models.

(1) a. Бесцветные зеленые идеи яростно спят.
b. Colorless green ideas sleep furiously.
(2) Я услышал ихний разговор.

В исследованиях по автоматической оценке приемлемости ставится задача обучить на массиве текстов языковую модель, которая будет оценивать приемлемость предложений наравне с человеком. Данные работы можно разделить на два типа: когнитивные и вычислительные. Первые из них, когнитивные, нацелены узнать, как устроена врожденная грамматика человека. Исследователи предполагают, что существует связь между вероятностью появления предложения в речи и его приемлемостью: чем чаще предложение встречается в корпусе, тем оно будет приемлемее. Другие же работы, вычислительные, изучают механизм обучения языковых моделей. Данные исследования пытаются ответить на вопрос, способны ли нейронные сети наравне с человеком определять приемлемость предложений, то есть обладают ли они, во-первых, аналогичных знанием грамматики, и, во-вторых, улавливают ли осмысленность и логичность высказываний.

Первой работой, в которой была осуществлена автоматическая оценка приемлемости, является статья Джея Лау, Александра Кларка и Шалома Лаппина [2016], которую можно отнести к когнитивному подходу. На национальном корпусе английского языка было обучено несколько языковых моделей: энграммная модель, скрытые марковские модели, рекуррентные нейронные сети. При обучении моделей были добавлены специальные метрики; предполагается, что они позволяют сгладить влияние побочных факторов, таких как лексическая частотность слов или длина предложения. Параллельно с этим при помощи краудсорсинга были собраны суждения о приемлемости 2500 предложений: из них 2000 предложений получены циклическим переводом 500 предложений из Британского национального корпуса на китайский, японский, норвежский, испанский с помощью Google Translate и обратно. Суждения о приемлемости предложений собирались по шкалам трех типов: 1-2, 1-4, 1-100. Далее языковые модели предсказывали вероятность 2500 предложений. Полученные меры вероятности, порожденные моделью, сравнивались со средними значениями оценок приемлемости носителей с помощью коэффициента корреляции Пирсона. Оказалось, что вероятностные модели демонстрируют довольно высокую точность в предсказании градуальных оценок приемлемости. Из этого Лау, Кларк и Лаппин делают вывод о градуальном устройстве грамматики.

Повторная попытка практического воплощения вероятностного подхода к грамматике была реализована в работе Джона Спрауза и коллег [2018]. Главное отличие данного исследования состоит в выборе предложений для вынесения суждений носителями и предсказания вероятностных метрик моделью. Предполагается, что использование предложений, полученных с помощью циклического машинного перевода, не совсем корректно, поскольку они не отражают более тонкие противопоставления, которые традиционно рассматриваются в экспериментальном синтаксисе. В исследовании Спрауза и коллег были использованы три различных датасета, каждый из который представляет интерес с точки зрения синтаксической теории. Первый набор данных состоял из минимальных пар из журнала Linguistic Inquiry, в которых изменению подвергался только один грамматический параметр. Во второй датасет были добавлены предложения из учебника по теоретическому синтаксису [Adger 2003]. Третья выборка содержала 120 предложений, которые получились путем перестановки слов во фразе (1b). В качестве языковых моделей были выбраны энграммная модель и рекуррентная нейронная сеть.

Результаты этого исследования дали результат, противоположный результату исследования Лау, Кларка и Лаппина. Спрауз и коллеги выяснили, что данные языковые модели не способны предсказать вариативность, которая предсказывается категориальными оценками в учебнике [Adger 2003], и не способны различить 23%-24% явлений, представленных в журнале Linguistic Inquiry. Следовательно, данная работа демонстрирует малую предсказательную силу вероятностного подхода.

Первым исследованием по автоматическому предсказанию оценок приемлемости в рамках вычислительного подхода является работа Алекса Варштадта, Аманприт Сингх и Самюэля Боумана [2019]. Вопрос, который стоял перед исследователями, звучал следующим образом: обладают ли предобученные нейронные модели знанием грамматики? В рамках данной работы был создан не просто датасет, а целый лингвистический корпус: он получил название Корпус лингвистической приемлемости (The Corpus of Linguistic Acceptability или CoLA). Он включает в себя больше 10 тысяч предложений на английском языке, взятых из литературы по теоретическому синтаксису. Корпус демонстрирует широкий разброс языковых феноменов как из области морфологии, так и из области синтаксиса. Предложения были размечены пятью экспертами-лингвистами по категориальной шкале: 0 (неприемлемо) или 1 (приемлемо). Варштадт, Сингх и Боуман применяли для предсказания оценок частично обученную модель с долгой краткосрочной памятью (LSTM). а также модели GPT, BERT, XLNet и T5, основанные на архитектуре трансформер. Модели обучались на британском национальном корпусе с использованием предобученных векторных представлений. При оценке качества моделей оказалось, что для модели LSTM значение коэффициента корреляции Мэтьюса (MCC) оказалось 0.341, для моделей BERT – 0.605, для модели GPT – 0.425, что сравнимо со значением 0.644 для оценок людей. Более того, значение MCC для модели XLNet оказалось выше, чем для оценок нелингвистов, а значение для T5 – выше, чем для оценок лингвистов.

Варштадт, Сингх и Боуман провели сравнение предсказательной силы моделей с архитектурой трансформер для разных языковых явлений. Оказалось, что наиболее высокое качество наблюдается для простых предложений. BERT и GPT преуспевают в предсказании оценок для предложений со сложной аргументной структурой, как, например, альтернации в глагольном управлении, а также для предложений со связыванием рефлексивных местоимений. Однако более сложные явления, такие как эллипсис, топикализация и вопросительные выносы, все еще сложно даются моделям. Важной отличительной чертой моделей GPT и BERT является тот факт, что они обучаются на огромном количестве предложений. В то время как одним из важнейших постулатов генеративного синтаксиса является аргумент о бедности стимула: ребенок усваивает язык благодаря врожденной грамматике, а не благодаря тому, что он сталкивался со всеми имеющимися в языке предложениями [Chomsky 1965]. Следовательно, данное исследование не является опровержением аргумента о бедности стимула и генеративной теории усвоения языка.

После публикации статьи Варштадта, Сингха и Боумана, задача автоматической оценки приемлемости стала приобретать все большую актуальность и была внесена в число задач GLUE (General Language Understanding Evaluation) [Wang et al. 2018]. GLUE представляет собой набор ресурсов для обучения, оценки и анализа систем понимания естественного языка. Конечной целью GLUE является стимулирование исследований в области разработки общих и надежных систем понимания естественного языка. Таким образом, включение задачи генерации оценок грамматичности в GLUE говорит о её актуальности и

важности в оценке работы нейронных языковых моделей. В GLUE также есть задачи, которые отвечают за понимание моделью смысла текста: например, логический вывод по тексту или анализ тональности.

Аналогичные корпуса лингвистической приемлемости, включающие предложения из лингвистических статей, размеченные по бинарной шкале, стали появляться и для других языков: ItaCoLA для итальянского [Trotta et al 2021], RuCoLA для русского [Михайлов и др. 2022]. Задача оценки приемлемости для русского языка была включена в датасет Russian SuperGLUE [Shavrina et al 2020], что несомненно подчеркивает ее актуальность. Тем не менее, корпуса наподобие CoLA имеют ряд проблемных мест. Предложения берутся из лингвистических работ, где не всегда рассматриваются однозначно приемлемые или неприемлемые языковые явления. Скорее наоборот, интерес лингвистов вызывают сложные неоднозначные феномены, для которых возникает вариативность в приемлемости между носителями. Этим можно объяснить тот неожиданный факт, что точность моделей XLNet и T5 даже превосходит человеческий результат на материале CoLA. Можно заметить, что человеческая оценка для данной задачи в целом оказывается не очень высокой: для русского языка на материале RuCoLA точность равна только 0.84, коэффициент корреляции Мэтьюса – 0.57. Авторы объясняют это тем, что оценка приемлемости непросто даётся и людям. Однако, на наш взгляд, дело именно в подборе материала, поскольку в данную задачу не следует включать предложения, для которых могут возникнуть несогласия между носителями.

Данная проблема была исправлена в более позднем исследовании Варштадта и коллег [2020]. В нем представлен корпус лингвистических минимальных пар (Benchmark of Linguistic Minimal Pairs, сокращенно BLiMP). Он состоит из 67 отдельных наборов данных, каждый из которых содержит 1000 минимальных пар, то есть пар минимально отличающихся предложений, которые различаются грамматической приемлемостью (3). В данном исследовании предложения были сгенерированы по специальным лингвистическим шаблонам. Это позволяет, во-первых, самостоятельно определять, какие именно феномены будут оценивать языковые модели, и, во-вторых, использовать минимальные пары. Для каждого корректного предложения на выбранное нами явление будет приходится одно некорректное, и различия между ними будут минимальны. Следовательно, можно говорить о том, что именно конкретная языковая ошибка влияет на автоматически предсказанную оценку. Важно также отметить, что приемлемость выбранных языковых феноменов не демонстрирует вариативности между носителями, а все ошибки, включенные в минимальные пары, имеют сугубо грамматическую природу. Таким образом, данное исследование нацелено именно на автоматическую оценку грамматичности.

(3) a. The cats annoy Tim. (*грамматично*)
b. *The cats annoys Tim. (*неграмматично*)
‘Кошки раздражают Тима’

Результаты исследования оказались следующими. Энграммная модель продемонстрировала точность 61.2%, рекуррентная сеть LSTM – 69.8, Transformer-XL – 69.6, GPT-2 – 81.5, тогда как точность для человеческой оценки равна 88.6. Было обнаружено, что современные модели надежно идентифицируют морфологические контрасты, связанные с согласованием, но более тонкие семантические и синтаксические явления, такими как элементы отрицательной полярности, эллипсис и вопросительные

выносы, вызывают трудности. Точность для нейронных сетей оказывается ниже, чем в исследовании 2019 года, однако точность для человеческой оценки наоборот повышается.

Исследования Варштадта и коллег на материале CoLA [2019] и BLiMP [2020] показывают, что для простых случаев вроде предложений с предикативным согласованием подлежащего и сказуемого точность оказывается довольно высокой, тогда как при дистантных вопросительных выносах она падает. Однако важно отметить, что предложения с предикативным согласованием и вопросительными выносами отличаются не только языковым феноменом, но и длиной: очевидно, более длинное предложение будет сложнее восприниматься как моделью, так и человеком. Вопрос состоит в том, будет ли предикативное согласование также легко усваиваться при наличии дистантных зависимостей.

Исследование Ребекки Марвин и Таля Линцена [2018] содержит дает ответ на данный вопрос. В данной работе осуществлялось автоматическое предсказание вероятности предложения от 0 до 1: соответственно, если вероятность близка к 0, предложение неграмматично, если близка к 1 – наоборот, грамматично. В данном исследовании рассматривалось три языковых явления: предикативное согласование, связывание рефлексива и единицы отрицательной полярности. Важно отметить, что относительно грамматичности данных феноменов не возникает вариативность среди носителей. Более того, предложения генерировались искусственным образом из ограниченного набора лексем, поэтому не всегда были семантически приемлемы. Следовательно, данное исследование решало именно задачу автоматической оценки грамматичности и было нацелено понять, насколько языковые модели усваивают грамматику.

Особенность работы состоит в том, что данные явления рассматривались в конфигурациях различной сложности: предложения с одним подлежащим и одним сказуемым (4a), с двумя сказуемыми (4b), сложноподчиненные предложения (4c), предложения с распространенным подлежащим (4d), предложения с относительной клаузой (4e) и т.д. Таким образом можно узнать, насколько хорошо модель различает, какое слово является подлежащим, какое – сказуемым, где проходит граница между главным и зависимым предложением. Для предсказания вероятности предложения на основе его грамматичности использовалась рекуррентная нейронная сеть LSTM. Результаты показали, что для простых предложений точность предсказания оказалась высокой, однако для более сложных предложений она значительно падает (4a-e).

(4) a. The author laughs/*laugh. (1.0)
‘Писатель смеется.’
b. The author laughs and swims/*swim. (0.90)
‘Писатель смеется и плывет.’
c. The mechanics said the author laughs/*laugh. (0.93)
‘Механик знает, что писатель смеется.’
d. The author next to the security guard smiles/*smile. (0.69)
‘Писатель рядом с охранником смеется.’
e. The author that knows the security guard laughs/*laugh. (0.74)
‘Писатель, который знает охранника, смеется.’

Подведем некоторые итоги. Сравнение предшествующих исследований показывает, что с появлением более мощных языковых моделей качество предсказания оценок приемлемости и оценок грамматичности растет. В связи с актуальностью данной задачи, появляются исследования с различными целями и задачами: как когнитивные, так и вычислительные; как с использованием предложений из лингвистических работ, так и с генерацией минимальных пар; как затрагивающие различные языковые явления, так и изучающие конкретные грамматические феномены. Для русского языка уже существует корпус RuCoLA, созданный для задачи оценки приемлемости на материале различных языковых явлений. Однако исследований, которые бы осуществляли автоматическую оценку грамматичности, нам не встречались.

В данной работе мы нацелены восполнить этот пробел и решить задачу автоматической оценки грамматичности для русского языка на материале предикативного согласования подлежащего и сказуемого. Выбор этого феномена обусловлен двумя причинами. Во-первых, согласование (операция Agree) является одной из базовых операций в синтаксисе после присоединения элементов для построения предложения (операция Merge) [Chomsky 1995]. В ходе согласования сказуемое копирует φ-признаки подлежащего (лицо, число) и приписывает ему именительный падеж. Во-вторых, предикативное согласование позволяет рассмотреть различные по сложности конструкции и выяснить, насколько хорошо языковая модель усваивает грамматику.

3. Датасеты

Поскольку языковая модель должна разграничить граммматичные и неграмматичные предложения, данные должны содержать не только корректные примеры, но и примеры с ошибками. В качестве датасета могут быть использованы искусственно сгенерированные предложения [Marvin & Linzen 2018; Warstadt et al. 2020]. 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. Наша цель состоит в автоматической оценке грамматичности: необходимо понять, насколько хорошо модель «понимает» устройство грамматики языка. Следовательно, некоторая неестественность предложений поможет абстрагироваться от лексических и прагматических аспектов и оценить именно знание грамматики.

Было сгенерировано 144 756 примеров. Мы исследовали, насколько хорошо модель способна усвоить механизм согласования по числу подлежащего и сказуемого: в грамматичных примерах число подлежащего и сказуемого будет одинаковым, в неграмматичных – разным. Таким образом, предложения датасета представляют собой минимальные пары: они отличаются ровно одним параметром, числом глагола. Это позволяет определить, насколько хорошо языковая модель усваивает целевое грамматическое явление. Данные содержат примеры различной степени сложности. Если в предложении имеется только одно существительное, подлежащее, и только один глагол, сказуемое (блок (a)), модель должна легко понимать, правильно ли указано число глагола. При наличии двух существительных, подлежащего и дополнения, предполагается, что модель может «запутаться» в согласовании при обратном порядке слов (блок (c)), тогда как при прямом порядке слов (блок (b)) трудностей не ожидается. Если в примере есть несколько существительных, то модели будет сложнее понять, грамматично ли

предложение. К примеру, мы ожидаем падение точности при распространенном подлежащем (блок (e)), в сложноподчиненном предложении (блок (d)), при наличии субъектного (блок (f)) и объектного (блок (g)) зависимого относительного предложения. Таким образом, всего датасет содержит 7 различных типов конструкций. Важно, что внутри каждого блока также есть разделение: половина предложений содержат одушевленные существительные, другая половина – неодушевленные. Это деление необходимо из-за того, что для неодушевленных существительных существует синкретизм (совпадение форм) именительного и винительного падежа, тогда как для одушевленных существительных эти формы различаются.

Данные включают в себя набор предложений, сгенерированных по определенным синтаксическим шаблонам из закрытого класса лексических единиц. Предложения представляют собой минимальные пары: корректный и некорректный примеры отличаются ровно по одному параметру, то есть некорректный пример содержит только одну ошибку. Ошибка отмечена знаком *. Шаблоны для конструирования предложений взяты из работы [Marvin & Linzen 2018] на материале английского языка и адаптированы для русскоязычного материала.

Набор лексических единиц, а также примеры предложений для каждого шаблона представлены ниже.

1. Подчинительный союз: *который, что*.
2. Подлежащее главного предложения (одушевленное): *писатель/ писатели, пилот/ пилоты, хирург/ хирурги, фермер/ фермеры, руководитель/ руководители, покупатель/ покупатели, командир/ командиры, преподаватель/ преподаватели, чиновник/ чиновники, студент/ студенты.*
3. Подлежащее главного предложения (неодушевленное): *фильм/ фильмы, журнал/ журналы, сериал/ сериалы, чертеж/ чертежи, учебник/ учебники, рисунок/ рисунки, роман/ романы, рассказ/ рассказы, спектакль/ спектакли.*
4. Подлежащее зависимого предложения: *охранник/охранники, ассистент/ ассистенты, архитектор /архитекторы, фигурист/ фигуристы, танцор/ танцоры, министр/ министры, водитель/ водители, секретарь/ секретари, офицер/ офицеры, родитель/ родители.*
5. Непереходный предикат главного предложения (одушевленный): *рассуждал/ рассуждали, пел/ пели, обедал/ обедали, был высоким/ были высокими, был пожилым/ были пожилыми, был молодым/ были молодыми, был низким/ были низкими.*
6. Непереходный предикат главного предложения (неодушевленный): *был хорошим/ были хорошими, был плохим/ были плохими, был новым/ были новыми, был популярным/ были популярными, был неизвестным/ были неизвестными*.
7. Переходный глагол: *знал/знали, ненавидел/ненавидели, недолюбливал/недолюбливали, презирал/презирали, ценил/ценили.*
8. Предлог: *рядом с, перед*.
9. Подлежащее зависимого предложения: *механик/ механики, банкир/ банкир.*
10. Глагол зависимого предложения: *говорил, думал, знал.*

(a) Простое предложение с подлежащим (532 примеров):
i. Единственное число (одушевленное):
Писатель рассуждал/ рассуждали*.

ii. Единственное число (неодушевленное):
*Фильм увлекал/*увлекали*.
iii. Множественное число (одушевленное):
Писатели рассуждали/ рассуждал*.
iv. Множественное число (неодушевленное):
Фильмы увлекали/ увлекал*.

(b) Простое предложение с подлежащим и дополнением, прямой порядок слов (7600 примеров):
i. Единственное / Множественное (одушевленное):
*Писатель знал/*знали охранников.*
ii. Единственное / Множественное (неодушевленное):
*Писатель знал/*знали рассказы*.
iii. Множественное / Единственное (одушевленное):
*Писатели знали/*знал охранника.*
iv. Множественное / Единственное (неодушевленное):
*Писатели знали/*знал рассказ*.
v. Единственное / Единственное (одушевленное):
*Писатель знал/*знали охранника*.
vi. Единственное / Единственное (неодушевленное):
*Писатель знал/*знали рассказ*.
vii. Множественное / Множественное (одушевленное):
*Писатели знали/*знал охранников.*
viii. Множественное / Множественное (неодушевленное):
*Писатели знали/*знал рассказы*.

(c) Простое предложение с подлежащим и одушевленным дополнением, обратный порядок слов (7600 примеров):
i. Множественное / Единственное (одушевленное):
*Охранников знал/*знали писатель*.
ii. Множественное / Единственное (неодушевленное):
*Рассказы знал/*знали писатель*.
iii. Единственное / Множественное (одушевленное):
*Охранника знали/*знал писатели.*
iv. Единственное / Множественное (неодушевленное):
*Рассказ знали/*знал писатели*.
v. Единственное / Единственное (одушевленное):
*Охранника знал/*знали писатель*.
vi. Единственное / Единственное (неодушевленное):
*Рассказ знал/*знали писатель*.
vii. Множественное / Множественное (одушевленное):
*Охранников знали/*знал писатели.*
viii. Множественное / Множественное (неодушевленное):
*Рассказы знали/*знал писатели*.

(d) Сложноподчиненное предложение (6384 примеров):

i. Множественное / Единственное (одушевленное):
*Механики говорили, что писатель рассуждал/*рассуждали.*
ii. Множественное / Единственное (неодушевленное):
*Механики говорили, что фильм увлекал/*увлекали.*
iii. Единственное / Единственное (одушевленное):
Механик говорил, что писатель рассуждал/ рассуждали.*
iv. Единственное / Единственное (неодушевленное):
*Механик говорил, что фильм увлекал/*увлекали.*
v. Множественное / Множественное (одушевленное):
*Механики говорили, что писатели рассуждали/*рассуждал.*
vi. Множественное / Множественное (неодушевленное):
*Механики говорили, что фильмы увлекали/*увлекал.*
vii. Единственное / Множественное (одушевленное):
*Механик говорил, что писатели рассуждали/*рассуждал.*
viii. Единственное / Множественное (неодушевленное):
*Механик говорил, что фильмы увлекали/*увлекал.*

(e) С зависимым существительным при подлежащем (16240 примеров):
i. Единственное / Множественное (одушевленное):
*Писатель рядом с охранниками был высоким/*были высокими.*
ii. Единственное / Множественное (неодушевленное):
*Фильм охранников был хорошим/*были хорошими.*
iii. Множественное / Единственное (одушевленное):
*Писатели рядом с охранником были высокими/*был высоким.*
iv. Множественное / Единственное (неодушевленное):
*Фильмы охранника были хорошими/*был хорошим.*
v. Единственное / Единственное (одушевленное):
*Писатель рядом с охранником был высоким/*были высокими.*
vi. Единственное / Единственное (неодушевленное):
*Фильм охранника был хорошим/*были хорошими.*
vii. Множественное / Множественное (одушевленное):
*Писатели рядом с охранниками были высокими/*был высоким.*
viii. Множественное / Множественное (неодушевленное):
*Фильмы охранников были хорошими/*был хорошим.*

(f) С субъектным зависимым предложением при подлежащем (53200 примеров):
i. Единственное / Множественное (одушевленное)
*Писатель, который знал охранников, рассуждал/*рассуждали.*
ii. Единственное / Множественное (неодушевленное)
*Писатель, который знал фильмы, рассуждал/*рассуждали.*
iii. Множественное / Единственное (одушевленное):
*Писатели, которые знали охранника, рассуждали/*рассуждал.*
iv. Множественное / Единственное (неодушевленное):
*Писатели, которые знали фильм, рассуждали/*рассуждал.*
v. Единственное / Единственное (одушевленное):
*Писатель, который знал охранника, рассуждал/*рассуждали.*

vi. Единственное / Единственное (неодушевленное):
*Писатель, который знал фильм, рассуждал/*рассуждали.*
vii. Множественное / Множественное (одушевленное):
*Писатели, которые знали охранников, рассуждали/*рассуждал.*
viii. Множественное / Множественное (неодушевленное):
*Писатели, которые знали фильмы, рассуждали/*рассуждал.*

(g) С объектным зависимым предложением при подлежащем (53200 примеров):
i. Множественное / Единственное (одушевленное):
*Писатели, которых знал охранник, рассуждали/*рассуждал.*
ii. Множественное / Единственное (неодушевленное):
*Фильмы, которые знал охранник, были хорошими/*был хорошим.*
iii. Единственное / Множественное (одушевленное):
*Писатель, которого знали охранники, рассуждал/*рассуждали.*
iv. Единственное / Множественное (неодушевленное):
*Фильм, который знали охранники, был хорошим/*были хорошими.*
v. Множественное / Множественное (одушевленное):
*Писатели, которых знали охранники, рассуждали/*рассуждал.*
vi. Множественное / Множественное (неодушевленное):
*Фильмы, которые знали охранники, были хорошими/*был хорошим.*
vii. Единственное / Единственное (одушевленное):
*Писатель, которого знал охранник, рассуждал/*рассуждали.*
viii. Единственное / Единственное (неодушевленное):
*Фильм, который знал охранник, был хорошим/*были хорошими.*

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

4. Train-test и кросс-валидация

Как было показано в предыдущем разделе, данные можно разделить на несколько групп, каждая из которых представляет собой отдельное языковое явление:

a) Простое предложение с подлежащим (532 примеров)
b) Простое предложение с подлежащим и дополнением, прямой порядок слов (7600 примеров)
c) Простое предложение с подлежащим и одушевленным дополнением, обратный порядок слов (7600 примеров)
d) Сложноподчиненное предложение (6384 примеров)
e) С зависимым существительным при подлежащем (16240 примеров)
f) С субъектным зависимым предложением при подлежащем (53200 примеров)
g) С объектным зависимым предложением при подлежащем (53200 примеров)

Следовательно, будет неправильно осуществлять рандомное разделение всего датасета. В этом случае есть вероятность, что в обучающей и в тестовой выборке языковые явления

будут представлены неравномерно. Тогда для тех примеров, которые были в тесте, точность будет выше.

Чтобы равномерно разделить датасет, мы применили следующую процедуру. Вначале каждая из 7 групп была перемешена, а затем поделена в следующем соотношении: 70% обучение, 15% валидация, 15% тест. После этого подчасти для обучающей выборки были собраны в одну обучающую выборку объемом 101327 предложений, подчасти валидационной – в одну валидационную объемом 21701, а подчасти тестовой – в одну тестовую объемом 21728 предложений, внутри выборок предложения также расположены в случайном порядке.

В данной работе использовалось трансферное обучение, то есть дообучение уже обученных языковых моделей на конкретную задачу. Поэтому данного количество предложений оказалось достаточно. В обучающей и тестовой выборке грамматичные предложения имели метку "1", предложения с ошибкой – метку "0". Следовательно, задача оценки грамматичности сводится к задаче классификации на два класса: класс грамматичных примеров с меткой "1" и класс неграмматичных с меткой "0".

5. Модели

В исследовании мы использовали трансферное обучение (transfer learning). Заранее обученные языковые модели с помощью полученного датасета дообучались на задачу классификации по грамматичности. В качестве моделей мы использовали как мультиязычные модели: BERT (104 языка) [Devlin et al. 2018], SlavicBERT (4 языка) [Arkhipov et al. 2019], XLM (15 языков) [Lample & Conneau 2019], так и модели для русского языка: ruBERT [Kuratov & Arkhipov 2019] и Russian RoBERTa [Blinov & Avetisian 2020]. Была осуществлена тонкая настройка (fine-tuning) для задачи классификации по грамматичности. Часть предложений из общего датасета были использованы для дообучения модели под конкретную задачу. Далее осуществлялось тестирование дообученных моделей на классификации по грамматичности. Данные были разделены в следующем соотношении: для трансферного обучения – 70%, для валидации – 15%, для тестирования – 15%. При обучении количество эпох было равно 5, использовался графический процессор.

В таблице 1 представлены результаты тестирования моделей после трансферного обучения. Можно заметить, что модели показывают отличное качество.

	Точность	Коэффициент корреляции Мэтьюса
ruBERT	1.0	1.0
Russian RoBERTa	1.0	1.0
Multilingual BERT	0.999	0.999
SlavicBERT	1.0	1.0
XLM Roberta	1.0	1.0

Таблица 1. Результаты тестирования предобученных языковых моделей в задаче классификации по грамматичности

Единственная модель, которая демонстрирует ошибки – Multilingual BERT. В таблице 2 представлена точность для модели по предложениям.

Тип предложения	Одушевленность существительного	Точность
Простое предложение с подлежащим	одушевленное	1.0
	неодушевленное	0.872
Простое предложение с подлежащим и дополнением, прямой порядок слов	одушевленное	1.0
	неодушевленное	1.0
Простое предложение с подлежащим и одушевленным дополнением, обратный порядок слов	одушевленное	1.0
	неодушевленное	0.998
Сложноподчиненное предложение	одушевленное	1.0
	неодушевленное	1.0
С зависимым существительным при подлежащем	одушевленное	1.0
	неодушевленное	1.0
С субъектным зависимым предложением при подлежащем	одушевленное	1.0
	неодушевленное	1.0
С объектным зависимым предложением при подлежащем	одушевленное	1.0
	неодушевленное	1.0

Таблица 2. Результаты для модели multilingual BERT при классификации различных типов предложений

Анализ результатов показывает, что наблюдается сложность в предсказании оценок грамматичности именно для неодушевленных существительных, что соответствует нашим ожиданиям. Следовательно, одушевленные и неодушевленные существительные по-разному воспринимаются моделью. Однако точность падает на предложениях с довольно простой структурой, что оказывается неожиданным. Можно сделать вывод, что сложность структуры не является значимой для моделей, обученных на больших объемах данных, какими являются трансформеры. В этом состоит отличие усвоения грамматики моделью и человеком. Мы предполагаем, что в данном случае точность зависит от количества обучающих данных в выборке: так, простых предложений было всего 532, предложений с прямым и с обратным порядком слов – по 7600 примеров, тогда как наиболее сложных предложений с субъектными и объектными относительными предложениями – по 53200 примера каждого типа. Тем не менее, интересно, что точность падает именно для предложений с обратным порядком, а не с прямым, что говорит о важности линейного порядка элементов для автоматической оценки грамматичности.

Таким образом, языковые модели на основе трансформеров показывают очень высокие результаты при дообучении на задачу оценки грамматичности. Усвоение правил предикативного согласования между подлежащим и сказуемым проходит успешно. Следовательно, все рассматриваемые модели обладают знанием грамматики и способны отличать существительные, различные по одушевленности. Из этого можно сделать вывод, что при обучении на больших массивах текстов трансформеры хорошо «выучивают» грамматику и усваивают синтаксическую структуру предложения вне зависимости от ее сложности.

6. Выводы

В данном разделе подведем выводы исследования.

Во введении (раздел 1) мы представили подробное определение понятия грамматичности и важности оценок грамматичности для теоретической лингвистики. Также мы рассмотрели, как суждения о грамматичности позволяют оценивать качество языковых моделей.

В обзоре предшествующих исследований (раздел 2) мы подробно остановились на работах, которые изучали автоматическое предсказание оценок приемлемости и оценок грамматичности. Мы описали разницу между приемлемостью и грамматичностью, постулируемую в лингвистике. Было показано, в чем различие когнитивного и вычислительного подхода к автоматической оценке приемлемости/ грамматичности. Мы рассмотрели данную задачу на материале разных языков и различных датасетов: как содержащих примеры из лингвистических исследований, так и включающие специально сгенерированные минимальные пары. Поскольку для русского языка уже существует корпус лингвистической приемлемости RuCoLA, в данном исследовании мы остановились на задаче автоматической оценки грамматичности для конкретного языкового явления: предикативного согласования подлежащего и сказуемого по числу. Данный феномен рассматривался на материале различных по сложности и по структуре предложений.

В частях работы, посвященных описанию данных (разделы 3 и 4), мы подробно представили процесс генерации данных по синтаксическим шаблонам из выбранного набора лексем. В ходе работы мы создали датасет из сгенерированных минимальных пар, размеченный по грамматичности. Также мы описали, как данные были поделены на обучающую (70%), валидационную (15%) и тестовую (15%) выборки.

При описании выбранных моделей и процесса обучения (раздел 5) был рассмотрен выбранный метод и набор моделей. Мы осуществили автоматическую оценку грамматичности для русского языка с использованием трансферного обучения. Языковые модели на основе трансформеров были дообучены на задачу классификации по грамматичнсти. Результаты оказались очень высокими: модели продемонстрировали значение меры точности и коэффициента корреляции Мэтьюса не менее 0.999. Предложения со сложной структурой не вызвали сложности в оценке грамматичности. Механизм внимания позволяет моделям осуществлять подробный синтаксический анализ, определять границы главного и зависимого предложения, различать синтаксические роли слов в предложении. Однако категория одушевленности вызывает небольшие затруднения: так, ошибки в оценке грамматичности возникают в предложениях с неодушевленными существительными и отсутствуют в предложениях с одушевленными существительными.

Таким образом, предобученные языковые модели на основе трансформеров действительно обладают знанием грамматики, а именно усваивают правила предикативного согласования подлежащего и сказуемого по числу вне зависимости от сложности синтаксической структуры предложения.

Данное исследование является законченной научной работой, которое планируется опубликовать в качестве научной статьи. Результаты автоматической оценки грамматичности являются важными для лингвистики и могут быть опубликованы в журналах «Вестник Московского университета. Серия 9. Филология» (RSCI, ВАК), «Вестник Санкт-Петербургского университета. Язык и литература» (RSCI, ВАК, SCOPUS, WOS), «Научный результат. Вопросы теоретической и прикладной лингвистики»

(SCOPUS), «Компьютерная лингвистика и интеллектуальные технологии» (SCOPUS), «Journal of Quantitative Linguistics» (SCOPUS, WOS). Кроме того, исследование представляет интерес для области обработки естественного языка и искусственного интеллекта, следовательно, может быть опубликовано в журналах «Интеллектуальные системы. Теория и приложения» (ВАК), «Научно-техническая информация. Серия 2: Информационные процессы и системы» (RSCI, ВАК), «Искусственный интеллект и принятие решений» (RSCI, ВАК), «Труды Института системного программирования РАН» (RSCI, ВАК), «Cybernetics and Information Technologies» (SCOPUS), «Lecture Notes in Computer Science» (SCOPUS).

Отметим, что исследование автоматической оценки грамматичности на материале русского языка может быть продолжено в последующих исследованиях на материале других конструкций: согласования по роду и лицу, падежного управления и вопросительных преобразований.

Литература.

Евгеньева, А. П. (1999) Словарь русского языка в 4-т. РАН, Ин-т лингвистических исследований. М.: Рус. яз., Полиграфресурсы.

Михайлов В., Шамардина Т., Рябинин М., Пестова А., Смуров И., Артемова Е. (2022). Насколько естественен естественный язык? Представляем датасет RuCoLA. Электронная публикация на сайте https://habr.com/.

Adger, D. (2003). *Core syntax: A minimalist approach* (Vol. 20). Oxford: Oxford University Press.

Arkhipov, M., Trofimova, M., Kuratov, Y., & Sorokin, A. (2019). Tuning multilingual transformers for language-specific named entity recognition. In Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing (pp. 89-93).

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

Chomsky, N. (1957). Syntactic structures. The Hague--Paris: Mouton. 1965 Aspects of the theory of syntax.

Chomsky, N. (1965). Aspects of the Theory of Syntax. Cambridge, MA: MIT Press.

Chomsky, N. (1995). *The Minimalist Program*, Cambridge, MA: MIT Press.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Kuratov, Y., & Arkhipov, M. (2019). Adaptation of deep bidirectional multilingual transformers for russian language. arXiv preprint arXiv:1905.07213.

Lample, G., & Conneau, A. (2019). Cross-lingual language model pretraining. arXiv preprint arXiv:1901.07291.

Lau, J. H., Clark, A., Lappin, S. (2016). Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge. Cognitive science, 41(5), 1202-1241.

Marvin, R., & Linzen, T. (2018). Targeted syntactic evaluation of language models. arXiv preprint arXiv:1808.09031.

Shavrina, T., Fenogenova, A., Emelyanov, A., Shevelev, D., Artemova, E., Malykh, V., Evlampiev, A. (2020). Russian SuperGLUE: A Russian language understanding evaluation benchmark. arXiv preprint arXiv:2010.15925.

Schütze, C. (2016). The empirical base of linguistics: Grammaticality judgments and linguistic methodology (p. 244). Language Science Press.

Sprouse, J., Yankama, B., Indurkhya, S., Fong, S., Berwick, R. C. (2018). Colorless green ideas do sleep furiously: gradient acceptability and the nature of the grammar. The Linguistic Review, 35(3), 575-599.

Trotta, D., Guarasci, R., Leonardelli, E., & Tonelli, S. (2021). Monolingual and cross-lingual acceptability judgments with the Italian CoLA corpus. arXiv preprint arXiv:2109.12053.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2018). GLUE: A multi-task benchmark and analysis platform for natural language understanding. arXiv preprint arXiv:1804.07461.

Warstadt, A., Singh, A., & Bowman, S. R. (2019). Neural network acceptability judgments. Transactions of the Association for Computational Linguistics, 7, 625-641.

Warstadt, A., Parrish, A., Liu, H., Mohananey, A., Peng, W., Wang, S. F., & Bowman, S. R. (2020). BLiMP: The benchmark of linguistic minimal pairs for English. Transactions of the Association for Computational Linguistics, 8, 377-392.