Научная статья

УДК 81’33
DOI 10.25205/1818-7935-2024-22-3-98-111

Бинарный классификатор для экспериментального поиска триггеров в шутках на английском языке

Евгения Максимовна Заковоротная

Национальный исследовательский университет «Высшая школа экономики»
Москва, Россия

haylin65@yandex.ru, https://orcid.org/0000-0002-3655-4369

Аннотация

Описывается создание модели, которая решает задачу распознавания юмористических и неюмористических текстов. Была обучена гибридная модель с предобученной нейронной сетью BERT в качестве эмбеддингового слоя и Bi-LSTM для классификации последовательностей. В качестве основного материала использовался обучающий и тестовый корпусы из 76 тысяч текстов, шуток и не-шуток. Особое внимание уделено идентичности лексики; данный критерий необходим, чтобы модель не распознавала разные категории текстов по лексике. В работе также описывается применение гибридной нейросети в серии экспериментов по лингвистическим преобразованиям юмористических и неюмористических текстов. Цель данных экспериментов заключается в поиске ключевых частей и слов, без которых шутка перестает быть юмористической. В рамках некоторых междисциплинарных теорий юмора подобные слова и выражения называют триггерами [Attardo S., 1994]. По результатам количественного и качественного анализа можно сделать вывод, что 78 из 100 шуток в валидационном датасете хотя бы один раз меняют метку класса на противоположную при использовании системы правил преобразований. При этом в 16 из оставшихся 22 шуток содержится явная или неявная экстралингвистическая информация. Т-критерий распределения Стьюдента, измеренный на вероятностных оценках исходного и измененного текста для каждого типа преобразования, позволил выявить преобразования, при которых чаще всего шутки из валидационного датасета перестают быть юмористическими: удаление панчлайна, удаление от 1 до 3 токенов с начала текста, удаление от 1 до 3 токенов с середины текста, удаление всех существительных.

Ключевые слова

автоматическая обработка естественного языка, изучение юмора, эксперименты, компьютерная лингвистика

Для цитирования

Заковоротная Е. М. Бинарный классификатор для экспериментального поиска триггеров в шутках на английском языке // Вестник НГУ. Серия: Лингвистика и межкультурная коммуникация. 2024. Т. 22, № 3. С. 98–111. DOI 10.25205/1818-7935-2024-22-3-98-111

Binary Classifier for Experimental Search of Triggers in Jokes in English

Eugeniia M. Zakovorotnaia

National Research University Higher School of Economics,
Moscow, Russian Federation

haylin65@yandex.ru, https://orcid.org/0000-0002-3655-4369

Abstract

This paper describes the development of a binary classifier differentiating humorous and non-humorous texts. The proposed seq2seq model consists of a pre-trained BERT embedding layer and a Bi-LSTM layer used for sequence classification. Training and validation corpora include 76,000 jokes and non-jokes with identical vocabulary; this is essential in preventing vocabulary choice from being employed as a distinguishing factor between humor and non-humor. Further, this paper also describes the application of the trained neural network in a series of experiments on linguistic transformations of humorous and non-humorous texts. The purpose of these experiments is to identify the essential parts and words, without which the joke ceases to be humorous. Some interdisciplinary theories of humor specify such expressions as triggers [Attardo S., 1994]. Based on the results of quantitative and qualitative analyses, 78 of the jokes from the validation dataset changed the label to the opposite at least once when the text was transformed. At the same time, 16 of the remaining 22 jokes contain explicit or implicit extralinguistic information. T-test, which measured probabilistic estimates of the original and modified texts for each type of linguistic transformation, revealed (keep tense consistent) the most common types of them: deletion of the punchline, deletion of the setup, deletion of 1 to 3 tokens from the beginning of the text, deletion of 1 to 3 tokens from the middle of the text and deletion of all the nouns.

Keywords

automatic natural language processing, humor study, experiments, computational linguistics

For citation

Zakovorotnaia E. M. Binary classifier for experimental search of triggers in jokes in English. *Vestnik NSU. Series: Linguistics and Intercultural Communication*, 2024, vol. 22, no. 3, pp. 98–111. DOI 10.25205/1818-7935-2024-22-3-98-111

Введение

Компьютерная лингвистика объединяет множество направлений междисциплинарных исследований, в том числе humor studies, где в качестве рабочих данных выступают тексты юмористического, иронического или саркастического характера. Источником подобного материала являются форумы и социальные сети, такие как Twitter, Reddit, Facebook и др. Собранные вместе, шутки составляют корпусы и коллекции данных, обладающие разметкой разной природы: например, юмористические тексты могут содержать оценку степени наличия комического эффекта [Blinov V. et al., 2019; Wang M. et al. 2020] либо оценку пригодности применения для решения различных задач [Hasan M. et al., 2020]. Исследовательские работы по юмору в основном описывают подходы к решению трех типов задач: классификация (определение, является ли текст шуткой), генерация (порождение новых шуток), создание корпусов (сбор, упорядочивание и разметка шуток), которые могут быть реализованы в рамках отдельных статей, либо участия в соревнованиях по изучению юмора, например, SemEval, 2021, SemEval2020, IberLEF2019. Отдельно стоит еще одна группа работ, направленная на изучение юмористических текстов с точки зрения формальных описаний либо уровней языка. Например, в работе [Karande A., 2006] авторы проводят практический разбор теоретических концепций того, насколько различные дискурсивные модели подходят для описания юмора. В статье [He H. et al., 2019] исследователи описывают подходы для измерения уровней семантического несоответствия в каламбурах и измененных с помощью омофонов текстах. В работах [Annamoradnejad I., 2022; Wang M. et al., 2020] описывается классификация шуток по наличию семантического несоответствия (сдвига) между частями шуток, а также обучение модели на разных частях шутки. В статье [Chen Y. et al., 2023] авторы использовали правиловую методологию Дж. Топлина

[Toplyn, J., 2023] для написания комедийных текстов в стиле “Late Night Show” и применения их в качестве промптов (prompt) на вход генеративной нейронной сети GPT-3. Данный подход позволил не только генерировать шутки с помощью GPT-3, но и регулировать тематику и стиль шуток. В работе [Tang L. et al., 2023] исследователи разработали классификатор юмористических текстов Naughtyformer на основе архитектуры Transformer, который хорошо распознает наличие оскорбительных паттернов в шутках.

Тем не менее, даже несмотря на использование инструментов искусственного интеллекта в данной области, среди исследователей теории юмора отсутствует формализация самой природы шуток, а также их объективных характеристик, которые могли бы быть подтверждены количественными методами. Кроме того, в большинстве работ по классификации юмористических текстов с точки зрения машинного обучения отсутствует использование языковых теоретических концепций о природе юмора. Лингвистическая составляющая в анализе юмора играет существенную роль: благодаря ей возможно отследить языковую игру, вычленить паттерны шуток, а также определить, насколько качественно работает классификатор или генератор шуток. Таким образом, лингвистический подход с использованием количественных методов становится актуальным при работе с юмористическими текстами.

В данной работе описывается обучение гибридного бинарного классификатора на основе BERT и Bi-LSTM, натренированного на распознавание шуток на английском языке. Также в статье описывается его применение для тестирования набора преобразований текста с помощью анализа изменения классификации текстов после удаления отдельных слов и целых фраз. Данная система представляет собой набор правил, в соответствии с которыми тексты изменяются с помощью удаления слов, словосочетаний, последовательностей. В измененном виде данные подаются на вход модели, которая классифицирует тексты как шутка или не-шутка.

Целью данной работы является обучение бинарного классификатора и поиск с его помощью таких фрагментарных преобразований текста, которые трансформируют юмористический текст в неюмористический. Мы считаем, что подобные слова и фразы обеспечивают связанность двух частей текста. В рамках семантической теории юмора [Morreall J., 2020; Raskin V., 1984; Raskin V., 2020] и общей теории вербального юмора подобные выражения называют триггерами.

Наши эксперименты позволят выяснить, какое влияние оказывает на восприятие конкретной шутки удаление ее отдельных частей. Полученные данные должны стать основой для правиловой системы преобразования текста, направленной на поиск ключевых слов и выражений. Ценность представленной работы заключается в том, что в отличие от работ [Weller O. Seppi K., 2019; Wang M., et al., 2020] (где совпадение лексики составляет не более 30 %) в собранных нами датасетах совпадение уникальной лексики юмористических и неюмористических текстов составляет более 80 %. Это предотвращает некачественное обучение классификатора, а именно устранение возможности распознавания типа текстов в зависимости их лексики.

Код с расчетами, экспериментами и данными хранится в репозитории на Github https://github.com/zijane/classifier_english_triggers.

1. Теоретические исследования

Семантическая теория и общая теория вербального юмора, которые анализируют проявление смешного через язык, базируются на философской теории несоответствия [Morreall J., 2020; Raskin V., 1984]. Согласно данному подходу, юмор основан на сочетании противоречивых концепций: добро и зло, день и ночь, светлое и темное и т. д.

Семантическая теория В. Раскина посвящена анализу вербального юмора и языковой игре. Она представляет каждую шутку с помощью структур знаний, фреймов или, иначе говоря, скриптов, которые выражают понятия, идеи, персонажи. С помощью таких конструкций

может быть формализован контекст и представление сюжета шутки. При описании разных сущностей, помимо однозначных слов, скрипты могут быть выражены одним или несколькими многозначными словами или выражениями. Если слово многоплановое, то формальное описание шутки должно отображать декомпозицию его значений выраженными более однозначными словами [Raskin V., 2020]. В. Раскин подчеркивает, что юмор возникает, только если оба фрейма находятся в оппозиции и одновременно семантически совместимы с одним и тем же текстом. На протяжении прочтения юмористического текста наблюдается семантический сдвиг, что представляет собой *смену одного семантического скрипта в пользу другого, иначе говоря, замещение первой концепции в пользу второй*. В шутках можно наблюдать разный уровень фреймовой оппозиции, где могут сочетаться исключительно противоположные идеи, либо альтернативы в рамках одной концепции, например, *хороший нрав vs плохой нрав* [Raskin V., 1984]. Семантический сдвиг обусловливает взаимное перекрытие (fully overlap) и частичное перекрытие (partial overlap) скриптов в коротких юмористических текстах. Кроме оппозиционных концепций, в реализации семантического сдвига участвует триггер – он напоминает о первом скрипте при прочтении второго, навязывает ему иную, неочевидную интерпретацию таким образом, чтобы сделать ее более правдоподобной и менее невозможной. В методологии семантической теории юмора выделяют два основных типа: двусмысленность (ambiguity) и противоречие (contradiction). Примечательно, что существует несколько вариантов триггера в зависимости от того типа двусмысленности, который присутствует в шутке. Противоречие, хотя и действует иначе, чем двусмысленность, создает также вторую интерпретацию, наложенную на весь текст, предшествующий триггеру, а также на текст, следующий за ним. Кроме того, иногда некоторые триггеры противоречия включают целые предложения, а не отдельные слова, и их связь с обычными триггерами противоречия аналогична связи между ситуационными неоднозначными триггерами и обычными неоднозначными триггерами. В. Раскин также упоминает случай, когда помимо основного триггера в шутке присутствует еще один вспомогательный – зачастую он нужен для того, чтобы ввести дополнительное значение, которое разнообразит описываемую в шутке ситуацию и усилит основной триггер.

Общая теория вербального юмора [Attardo S., 1994] С. Аттардо и В. Раскина сочетает в себе не только элементы семантики, но и текстовой лингвистики, теории нарратива и прагматики. Методология рассматривает не только оппозицию фреймов, но вводит дополнительные критерии, источники знаний (Knowledge Resources) для анализа шутки: логический механизм, цель, стратегия повествования, язык и ситуация. Триггер семантической теории заменяется здесь на логический механизм, который объясняет, каким образом два смысла (скрипты) объединяются в шутке, и как при их прочтении реализуется семантический сдвиг. Данное понятие включает не только двусмысленность и противоречие, но и ложные аналогии, к примеру Garden-Path phenomena [Pritchett Bradley L., 1988; Raskin V., 1984] или текстовые figure-ground reversals [Raskin V., 2020; Veale T., 2008].

Здесь мы можем высказать гипотезу о том, что устранение триггера должно приводить к тому, что шутка перестает быть таковой. Это может быть справедливо не для всех случаев, но изменения должны быть статистически значимыми.

Для того чтобы автоматизировать проверку нашей гипотезы, мы должны обучить нейросетевой классификатор, который будет с высокой точностью различать юмористические и неюмористические тексты. Устранение триггера или другого элемента, влияющего на восприятие текста как юмористического, должно снижать степень уверенности классификатора в том, что перед ним шутка.

Заметим, что подобные эксперименты не будут отличать изменения, связанные с удалением собственно триггера, от изменений, связанных с устранением слов, важных для понимания сути шутки. Более того, возможна и обратная ситуация, при которой устранение части слов из неюмористического текста приведет к тому, что он станет восприниматься классификатором как юмористический. В связи с этим требуется анализ изменений, произведенных над текстом.

2. Данные

Прежде чем перейти к практической части, заключающейся в разработке классификатора и качественном анализе результатов, важно описать рабочий материал, на котором проводились эксперименты. Основным корпусом в данном исследовании является набор в 76 100 фрагментов текстов. Он состоит из подкорпусов шуток и не-шуток примерно одинакового объема. Подкорпус шуток насчитывает 38 100 штук на английском языке. Он был составлен из более чем 20 датасетов, найденных на сайтах Kaggle и Github. Минимальное количество слов в шутках 3, максимальное – 32, среднее – 12. Количество уникальных лемм в шутках – 9 931. Размер подкорпуса шуток в словоупотреблениях – 472 475.

Во время экспериментов шутки могли быть разделены на сетап и панчлайн. В терминах вышеупомянутых междисциплинарных теорий юмора сетап служит для определения первого скрипта, в котором задается исходная ситуация. Панчлайном характеризуется часть, где первичная интерпретация шутки изменяется с помощью уточнения второго скрипта. В данном исследовании панчлайн всегда представлен последним предложением либо конечной частью предложения после запятой; все остальное считается сетапом.

Для сравнения с шутками использовался неюмористический датасет объемом в 38 000 фрагментов текстов. Он был составлен на основе 10 корпусов новостей, заголовков и текстов, а также на основе 600 художественных произведений (романов, повестей, рассказов, часть из которых относится к юмористическим) на английском языке. Эта часть текстов была найдена на сайте Kaggle и в электронной библиотеке Guttenberg. Использование большого количества источников обусловлено целью подобрать для юмористических текстов максимально схожие по лексике фрагменты неюмористических. Это необходимо для того, чтобы используемая модель при обучении не использовала конкретную лексику в определенных типах текстов как подсказку. Данный этап при подготовке корпуса возник после тестирования первичной модели, обученной на текстах с отличающейся лексикой в шутках и не-шутках. Первичные результаты тренировки показывали на валидационном датасете Accuracy – 0,988, что на 2 % выше показателя у финальной версии классификатора. На тестовом датасете предварительная модель сработала хуже и выдала Accuracy – 0,91, а итоговая – 0,93. Минимальное количество слов в не-шутках – 2, максмальное – 28, среднее – 8. Количество уникальных слов в не-шутках – 11 306. Размер подкорпуса не-шуток в словоупотреблениях – 310 009.

При составлении финальной версии общего подкорпуса не-шуток сравнивался с подкорпусом шуток на предмет схожести лексики. Если леммы в конкретной не-шутке совпадали с леммами 38 100 шуток, то тогда не-шутка входила в общий корпус. Задача состояла в том, чтобы добиться максимального совпадения лемм в обоих подкорпусах.

Распределение частей речи в шутках и не-шутках представлено в табл. 1.

Таблица 1

Распределение частей речи в шутках

Table 1

POS-tags distribution in jokes.

Частеречный тег	**Количество лемм с конкретным тегом шутки**	**Количество лемм с конкретным тегом не-шутки**	**Количество общих лемм на каждый тег**
1	**2**	**3**	**4**
NOUN	4639	5151	3310
ADJ	1612	1898	1084
PROPN	1431	1687	777
VERB	1401	1651	847
ADV	388	484	296

Окончание табл. 1

1	2	3	4
NUM	195	207	154
PRON	66	52	47
ADP	55	62	46
AUX	42	34	20
INTJ	41	21	11
SCONJ	24	24	20
X	16	7	2
DET	11	19	10
CCONJ	6	7	5
PART	2	1	1
SPACE	1	0	1
SYM	1	1	1

Пересечение лемм для юмористического и неюмористического корпуса составляет 8 361 уникальное слово.

В данном разделе описывается обучение гибридного бинарного классификатора на решение задачи распознавания текстов по типу юмор/не-юмор. После того как датасет с двумя типами текстов был подготовлен, он был разделен на три части: для тренировки (57 000 шт.), для тестов (19 000 шт.), для валидации (200 шт. – 100 шуток и 100 не-шуток). Для обучения классификатора применялись только два датасета: тренировочный и тестовый. Валидационный датасет был составлен из шуток, длина которых была не менее 4 слов. В противном случае удаление слов из шутки могло дать вырожденный текст.

3. Архитектура нейросетевого классификатора

В данной работе использовалась гибридная модель, состоящая из слоев предобученной нейронной сети с архитектурой Transformer BERT [Devlin J. et al., 2019] и двунаправленной LSTM-модели для обработки последовательностей (BiLSTM). В качестве Embedding Layer для конструирования контекстно-зависимых векторных представлений текстов применялся BERT, а именно слои модели “bert-based-uncased” [Devlin J. et al., 2019]. Это англоязычная языковая модель, предварительно натренированная на решение задачи предсказания замаскированного токена (masked language modeling). Данная модель не видит разницы между нижним и верхним регистром. Нейронная сеть (Bi-LSTM) использовалась в качестве классификатора, принимая на вход скрытые представления последовательностей, созданные на предыдущем слое Embedding Layer. Характеристики модели представлены в табл. 2.

Таблица 2

Описательные характеристики нейросетевого бинарного классификатора

Table 2

Descriptive characteristics of a neural binary classifier.

Режим обучения модели	with torch.no_grad()
Длительность обучения модели	5 эпох
Количество units в LSTM()	64
Тип свертки	tf.keras.layers.GlobalAveragePooling1D(), tf.keras.layers.GlobalMaxPooling1D()

Окончание табл. 2

Лейблы	[0,1], где 0 – это “joke”, а 1 – это “non-joke”
Коэффициент на слое Dropout()	0,3
Значение на слое Dense()	2
Функция активации	softmax
Оптимизатор	tf.keras.optimizers.Adam()
Функция потерь	Бинарная кросс-энтропия

При анализе результатов и составлении экспериментов проводилась морфосинтаксическая разметка юмористических текстов при помощи англоязычной модели ‘en_core_web_sm’ библиотеки Spacy для языка Python. Она представляет собой сверточную нейронную сеть, которая умеет решать задачи частеречной и морфосинтаксической разметки, а также распознавания именованных сущностей через пайплайн и компоненты Spacy [https://spacy.io/models/en#en_core_web_sm].

Важно отметить, что по завершении тренировки модели для получения контрольных значений валидационные данные были поданы на вход модели без изменений. Accuracy – 0,965; F1 – 0,964; Precision – 0,967; Recall – 0,963.

4. Эксперименты

В данном разделе описывается проведение серии экспериментов влияния лингвистических трансформаций на автоматическую обработку юмористических текстов. В ходе отдельного эксперимента мы удаляли заданные текстовые фрагменты, являющиеся потенциально значимыми смысловыми концептами, из всех текстов. Данный подход является релевантным, скорее, для шуток, чем для новостей или частей художественных произведений, поскольку последние изначально являются обычными, неюмористическими текстами [Raskin V., 1984]. Результаты экспериментов приведены для части валидационного корпуса, где было представлено 100 шуток. После того как были получены первичные оценки модели, валидационный датасет подвергся серии экспериментальных преобразований с целью выявить, при каких изменениях шутка перестает быть юмористической и превращается в обычный текст.

Удаление фрагментов из текстов мотивируется их потенциальным влиянием на смысловые значения в разных частях текстов, особенно в шутках. Рабочей единицей является токен – слово или знак препинания. При построении экспериментов учитывалась позиция токенов относительно структуры последовательности, их лексическое выражение, а также синтаксические связи.

1. Удаление токенов с конца текста. Данный эксперимент исходит из определения панчлайна шутки, как части, в которой происходит смена скриптов и возникает юмористический эффект. На вход модели поступали последовательности с удаленными последними токенами (от 1 до 3 шт.). Количество обусловлено минимальной длиной юмористических текстов 3.

2. Удаление с начала текста. Данный эксперимент опирается на определение сетапа шутки как части, где вводится первый скрипт и выражается первичное значение юмористического текста. На вход модели поступали последовательности с удаленными первыми токенами (от 1 до 3 шт.). Количество обусловлено минимальной длиной юмористических текстов 2. В случае, когда количество удаляемых слов превышало исходную длину шутки, то процесс удаления токенов из юмористического текста не производился.

3. Удаление из середины текста. Данный эксперимент опирается на структуру юмористического текста, где в середине происходит смена сетапа на панчлайн. В данном случае середина рассчитывалась как половина длины текста с учетом пунктуации, и тогда слово, находяще-

еся на этой позиции, и являлось серединой. На вход модели поступали последовательности, где посередине удалялось от 1 до 3 токенов. Количество обусловлено минимальной длиной юмористических текстов.

4. Удаление одного существительного/имени собственного/местоимения. Данный эксперимент исходит из грамматических характеристик указанных частей речи, которые определяют существительное как обозначение одушевленного и неодушевленного предмета, понятия, идеи, концепции, а местоимение – указание или замещение предмета, качества или числа в зависимости от части речи. В существительном, а также местоимении содержится основная доля семантического контекста, задаваемого предложением в целом. При удалении данных частей речи контекст должен измениться и стать более размытым. Следовательно, при удалении слова, важного для повествования, шутка должна перестать быть шуткой, а модель должна присвоить измененному тексту другой лейбл. На вход модели поступали последовательности трех типов: с удаленным попеременно в сетапе, панчлайне и в целом тексте любым первым токеном сначала, у которого часть речи определяется либо как существительное (NOUN), либо имя собственное (PROPN), либо местоимение (PRON). Выбор удаляемого слова определяли не части речи, а порядок слов в предложении.

5. Удаление всех существительных/имен собственных/местоимений. Идея и предпосылки совпадают с основой предыдущего эксперимента. В данном случае из последовательности извлекаются все существительные для того, чтобы впоследствии сравнить влияние на контекст последовательности одного случайного существительного/местоимения/имени собственного и всех, которые есть. На вход модели поступали последовательности с удаленным попеременно в сетапе, панчлайне, и целом тексте токенами, у которых часть речи определяется как существительное (NOUN), имя собственное (PROPN), местоимение (PRON).

6. Удаление одного глагола. Данный эксперимент исходит из грамматических характеристик указанной части речи, которые определяют глагол как обозначение состояния или действия предмета. Предикат является основой синтаксического дерева и обладает связью как с главными, так и со служебными частями речи. Вспомогательные глаголы уточняют и дополняют семантическое значения связанной части речи. При удалении данной части речи контекст должен измениться и стать более размытым. Шутка может перестать быть шуткой, и модель может присвоить измененному тексту другой лейбл. На вход модели поступали последовательности с удаленным попеременно в сетапе, панчлайне и целом тексте одним токеном, у которого часть речи определяется как глагол (VERB) или вспомогательный глагол (AUX).

7. Удаление всех глаголов. Идея и предпосылки совпадают с предыдущим экспериментом. Однако здесь исключаются все слова заявленной части речи из последовательности, чтобы впоследствии провести сравнительный анализ о влиянии глагола на контекст исходной последовательности. На вход модели поступали последовательности с удаленными попеременно в сетапе, панчлайне и целом тексте токенами, у которых часть речи определяется как глагол (VERB) или вспомогательный глагол (AUX).

8. Удаление одной именной группы, элементы которой объединены синтаксическими связями nsubj (именное подлежащее), expl (плеоназные номинации), nmod (атрибут или дополнение в родительном падеже), appos (зависимое нарицательное для модификации первого существительного). Данный эксперимент объединяет случаи, когда существительное в основной или зависимой ролях выражает главное или дополнительное значения. На вход модели поступали последовательности с удаленными попеременно в сетапе, панчлайне, и целом тексте токенами, между которыми тип синтаксической связи является одним из вышеуказанных.

9. Удаление одного словосочетания, элементы которого объединены синтаксическими связями nsubj (именное подлежащее). Данный эксперимент объединяет случаи, когда существительное в основной роли и выражает главное значение всего или части текста. На вход мо-

дели поступали последовательности с удаленными попеременно в сетапе, панчлайне и целом тексте токенами, между которыми тип синтаксической связи является nsubj.

10. Удаление одного словосочетания, элементы которого объединены синтаксическими связями типов expl (плеоназные номинации), nmod (атрибут или дополнение в родительном падеже), appos (зависимое нарицательное для модификации первого существительного). Данный эксперимент объединяет случаи, когда существительное в основной или зависимой ролях и выражает дополнительные значения. На вход модели поступали последовательности с удаленными попеременно в сетапе, панчлайне и целом тексте токенами, между которыми тип синтаксической связи является одним из вышеуказанных.

11. Удаление отдельных частей шутки. Данный эксперимент предполагает радикальное преобразование для исходного текста, особенно для юмористического, так как удаляется одна из двух частей текста. Это применяется для выявления ключевой части шутки, а также для проверки такой гипотезы, что при удалении панчлайна все шутки перестанут быть юмористическими и будут восприниматься моделью как обычный текст. Кроме того, проводится как совместное удаление сетапа/панчлайна в рамках одного эксперимента, так и отдельное в рамках двух экспериментов.

При валидации работы модели и формировании количественных оценок в каждом эксперименте были взяты наименьшие показатели вероятностей, с которыми были определены лейблы. Данные были объединены в единую таблицу, где повторно были выделены наименьшие показатели вероятности оценки текста для каждой из последовательности. На основе данной таблицы была измерена статистическая метрика Accuracy. В рамках количественного анализа результатов было также проведено измерение t-критерия Стьюдента для измерения двух выборок: между начальными вероятностными оценками присвоения лейбла по текстам и наименьшими вероятностными оценками присвоения лейбла, полученными во время каждого эксперимента. В табл. 3 содержатся оценочные показатели, измеренные только на юмористических текстах из валидационного датасета.

Таблица 3

Оценочные показатели результатов экспериментов
для шуток из валидационного датасета

Table 3

Estimated experimental results for jokes from the validation dataset

Виды преобразований	**Accuracy**	**p-value**
Текст без изменений	0,93	–
Удаление от 1 до 3 токенов с конца текста	0,77	< 0.001
Удаление от 1 до 3 токенов с начала текста	0,84	< 1e-4
Удаление от 1 до 3 токенов в середине текста	0,91	< 0,001
Удаление одного токена типа NOUN/PROPN/PRON из текста	0,91	< 0,001
Удаление всех слов типа NOUN/PROPN/PRON из текста	0,73	< 1e-4
Удаление одного токена типа VERB/AUX из текста	0,89	< 0,001
Удаление всех токенов типа VERB/AUX из текста	0,88	< 1e-4
Удаление одного словосочетания типа nsubj +nmod, expl, appos из текста	0,87	< 0,005
Удаление одного словосочетания типа nsubj из текста	0,87	< 0,005
Удаление одного словосочетания типа nmod, expl, appos из текста	0,94	> 0,05
Удаление только сетапа из текстов	0,84	> 0,05
Удаление только панчлайна из текстов	0,51	< 1e-4
Удаление одного сетапа/панчлайна из текста	0,55	< 1e-4

6. Результаты

В данном разделе описывается проведение качественного анализа данных экспериментов по удалению фрагментов текстов, а также рассчитываются два дополнительных показателя, которые позволяют лучше охарактеризовать итоговую информацию. Кроме того, проводится анализ показателей p-value, рассчитанных только для шуток из валидационного датасета.

Первый показатель отражает количество экспериментов, где лейбл, определенный моделью, не совпадает с исходным, корректным лейблом. Следует отметить, что в расчетах также учитывался лейбл, проставленный моделью тексту без преобразований. Сравнительный анализ данных, которые были получены по итогам проведения всех экспериментов, позволил выявить, что 78 шуток из всего валидационного датасета хотя бы один раз меняют лейбл на противоположный. Остальные 22 шутки остаются юмористическими несмотря на все изменения последовательностей.

Качественный анализ 22 шуток, у которых лейбл ни разу не менялся на противоположный, позволил выявить, что они обладают ярко выраженными общими чертами. С точки зрения синтаксиса, в основном это шутки-загадки (вопрос-ответ) (65 %), при этом также есть шутки в одно-два предложения (34 %), один диалог. Шутки нарративного формата отсутствуют. При этом 16 из 22 шуток содержат явную или неявную экстралингвистическую информацию. Экстралингвистика, или внешняя лингвистика, – раздел языкознания, исследующий связь языка с мышлением, обществом, культурой, действительностью [Epstein B., 2008]. Так, большинство примеров из ограниченной выборки содержат упоминание национальностей, религиозных групп, технологий, продуктов, известных политических и общественных деятелей. В остальных 25 % обыгрывается ситуативная двусмысленность. Кроме того, важно отметить, что 15 из 22 шуток, где ни разу не менялся лейбл, содержат вероятностные оценки модели выше порога 0,7.

Качественный анализ шуток, где менялся лейбл хотя бы раз, продемонстрировал, что их характеристики схожи с предыдущей группой текстов. Например, основной синтаксической конструкцией является «вопрос-ответ» (57 %), вторичной – тексты в одно-два предложения (43 %). Нарративные и диалоговые шутки отсутствуют в анализируемых данных. Элементы внешней лингвистики, именованные сущности, встречаются в 22 текстах данной группы. Преобладают имена знаменитостей и этнические группы.

Проверка именованных сущностей на предмет их наличия в новостных и художественных отрывках, использованных при обучении классификатора, показала, что в обеих группах шуток не нашлось по два имени знаменитостей, в то время как все остальные слова встретились. Это подтверждает, что обученный классификатор, в основе которого лежит BERT-модель, при распознавании текстов ориентируется не на лексику, а на синтаксис или семантику.

Необходимо обратить внимание на табл. 3, в которой отображены результаты измерения t-критерия Стьюдента на первичных вероятностных оценках модели и минимальных вероятностных оценках, полученных при разных преобразованиях текстов шуток. Показатели демонстрируют, что наиболее незначительными для изменения юмористических текстов являются удаления всех именных групп и удаления одного словосочетания типа nmod, expl, appos из текста. Наиболее влиятельными преобразованиями текста на вероятностные оценки модели стали различные виды удаления слов, среди которых лидирующую позицию занимает попеременное удаление панчлайна из текстов. Если разделять удаление скриптов в юмористических текстах, то наиболее влиятельным на текст оказывается удаление панчлайна из юмористических текстов. Это подтверждает, что в большинстве шуток из валидационной выборки смена скриптов и возникновение двусмысленности описываемых концепций, а также наличие триггеров заключено все-таки во второй половине последовательности. Примечательно, что удаление слов из начала юмористического текста влияет сильнее на вероятностные оценки модели, чем удаление от слов с конца. Такой результат, полученный на данном датасете, можно объяснить

наиболее часто встречающимися синтаксическими связями между словами в начале и конце юмористического текста. В табл. 4 и 5 показано количество синтаксических связей между первыми тремя токенами и последними тремя токенами соответственно. Синтаксические связи были выявлены с помощью англоязычной модели синтаксического анализа spaCy. Табл. 4 демонстрирует, что наиболее распространенной синтаксической связью в первых трех токенах с начала шутки является именное подлежащее nsubj, которое вместе со сказуемым выражает основную идею последовательности. Табл. 5 показывает, что наиболее частые синтаксические связи между последними тремя токенами, помимо пунктуации, выражены сказуемым (root), прямым дополнением (dobj), предложным дополнением (pobj), которые дополняют, но не выражают основную концепцию подлежащего.

Таблица 4

Частота встречаемости синтаксических связей
между первыми тремя токенами с начала шутки

Table 4

Frequency of occurrence of syntactic relations between the first three tokens
from the beginning of the joke

Синтаксические зависимости по spaCy	**Встречаемость**
nsubj	66
aux	45
ROOT	45
advmod	36
det	24
dobj	18
prep	13
compound	10
amod	8
прочие виды связей	35

Таблица 5

Частота встречаемости синтаксических связей
между первыми тремя токенами с конца шутки

Table 5

Frequency of occurrence of syntactic connections between the first three tokens
from the end of the joke

Синтаксические зависимости по spaCy	**Встречаемость**
punct	116
ROOT	43
dobj	22
pobj	21
det	12
prep	10
amod	9
attr	8
advmod	8
compound	7
прочие виды связей	43
nsubj	1

Если сравнить показатели t-теста между удалением от 1 до 3 токенов с начала последовательности и удалением сетапа, то разница оказывается значительной. Удаление ключевых частей речи NOUN/PRON/PROPN, которые заключают в себе основные смысловые концепты и идеи текста, оказывает равноценное предыдущему типу влияние на вероятностные оценки модели. При этом удаление всех видов глагола не так сильно искажает смысл текста, а соответственно и принятие решения моделью. Однако удаление одного существительного и одного глагола обладает почти одинаковым влиянием на модель.

Кроме того, важно отметить, анализ показал, что только 18 из 100 шуток получили минимальную вероятностную оценку исключительно через удаление сетапа-панчлайна. При этом на 40 % из них это преобразование никак не повлияло в изменении лейбла. То есть при исключении данного подхода все остальные методы удаления слов и словосочетаний можно использовать в качестве формальных правил для тренировки механизмов внимания будущего распознавателя ключевых слов шутки.

Заключение

Данная работа описывает разработку нейронного бинарного классификатора, распознающего юмористические и неюмористические тексты. С помощью модели была протестирована система формальных подходов по лингвистическим преобразованиям текстов с помощью удаления слов и словосочетаний.

Прежде всего, наш подход к построению и обучению классификатора отличается от подходов, применяемых в предыдущих работах. Предварительные тесты продемонстрировали, что если не отфильтровать юмористические и неюмористические тексты по лексике, чтобы она была схожей, если не одинаковой, то классификатор обучается распознавать тексты не по семантике или синтаксису, а по отдельным словам, которые встречаются только в конкретных типах текстов. В результате нами была обучена гибридная модель с BERT в качестве эмбеддингового слоя на корпусе из 76 100 фрагментов текстов, шуток и не-шуток, с идентичной лексикой. Accuracy и F1-мера составили 0,965 и 0,964 соответственно. Кроме того, выбранный нами подход опирается на теоретические работы В. Раскина и С. Аттардо по юмору в области лингвистики в плане составления экспериментов и поиска ключевых слов в шутках.

С использованием модели было протестировано влияние различных типов преобразований юмористических текстов с целью поиска ключевых частей и слов, без которых шутка перестает распознаваться как юмор. Сравнительный анализ данных, которые были получены по итогам проведения всех экспериментов, позволил выявить, что 78 из 100 шуток из всего того датасета хотя бы один раз меняют лейбл на противоположный. Остальные 22 шутки остаются юмористическими несмотря на все изменения последовательностей. Качественный анализ продемонстрировал, что они обладают ярко выраженными общими чертами. С точки зрения синтаксиса, в основном это шутки-загадки (вопрос-ответ). При этом 16 из 22 шуток содержат явную или неявную экстралингвистическую информацию, а в остальных 6 обыгрывается ситуативная двусмысленность.

В дальнейшем планируется применить полученную модель на расширенном корпусе шуток для определения доли подобных исключений. При этом важно использовать несколько датасетов разного размера: 100 шт., 1000 шт, 5000 шт., 10000 шт. – главное, чтобы юмористические тексты не встречались в тренировочном датасете. Данная операция позволит определить, была ли полученная доля исключительных шуток случайной или закономерной. Кроме того, это позволит выявить дополнительные характеристики исключительных шуток и определить их формальные паттерны.

Т-критерий распределения Стьюдента, измеренный на вероятностных оценках исходного и измененного текста для каждого типа преобразования, позволил выявить самые влиятельные

типы преобразования текста: удаление панчлайна, удаление от 1 до 3 токенов с начала текста, удаление всех существительных, удаление глаголов.

Полученные результаты и метрики на данном этапе демонстрируют потенциал в представленном подходе обучения классификатора на распознавание юмора. Эксперименты с лингвистическими преобразованиями подтвердили гипотезу о том, что существительные, а также местоимения и имена собственные являются носителями основного контекста. Следовательно, при поиске триггеров в шутках важно обращать внимание на них в первую очередь. Однако паттерны с удалением первого существительного/местоимения/имени собственного/глагола показали свою несостоятельность по сравнению с остальными экспериментами. Это означает, что основной контекст может быть сосредоточен в разных частях шутки, и требуется иной подход выявлению ключевых существительных. Необходимо также обращать внимание на позиции слов в юмористических последовательностях – хотя удаление панчлайна обладает наибольшим влиянием на смену лейбла в шутке из представленных экспериментов, удаление слов из начала юмористического текста влияет сильнее на вероятностные оценки модели, чем удаление от слов с конца. Это объясняется присутствующими синтаксическими связями между словами в начале и конце юмористических текстов из валидационной выборки. В начале шуток в пределах первых трех слов зачастую присутствует номинальный субъект (nsubj), а в конце в пределах трех слов – различные виды дополнения (dobj, pobj и т.д.). Таким образом, на данном датасете подтверждается заявленная структура шутки, где в первом скрипте вводятся основные персонажи, события, понятия, концепты, а во втором заявленные образы дополняются иными идеями.

Список литературы / References

Annamoradnejad I. ColBERT: Using BERT Sentence Embedding for Humor Detection. 2022, URL: https://arxiv.org/abs/2004.12765

Attardo S. Linguistic theories of humor. Mouton de Gruyter. 1994

Blinov V., Bolotova-Baranova V., Braslavski P. Large Dataset and Language Model Fun-Tuning for Humor Recognition // In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, p. 4027–4032.

Chen Y., Shi B., Si M. Prompt to GPT-3: Step-by-Step Thinking Instructions for Humor Generation. 2023, URL: https://arxiv.org/abs/2306.13195

Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019, vol. 1 (Long and Short Papers), p. 4171–4186.

Epstein B. The Internal and the External in Linguistic Explanation. // Croatian Journal of Philosophy, 2008, vol. 8(22), p. 77–111.

Hasan M. K., Rahman W., Zadeh A. B., Zhong J., Tanveer M. I., Morency L.-P., Hoque M. UR-FUNNY: A Multimodal Language Dataset for Understanding Humor. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, p. 2046–2056.

He H., Peng N., Liang P. Pun Generation with Surprise // In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019, p. 1734–1744.

IberLEF2019, URL: https://sites.google.com/view/iberlef-2019/

Karande A. What Humour Tells Us About Discourse Theories // Conference of the European Chapter of the Association for Computational Linguistics, 2006, p. 31–38.

Liu Y., Ott M., Goyal N., Du J., Joshi M, Chen D., Levy O., Lewis M., Zettlemoyer L., Stoyanov V. // 'RoBERTa: A Robustly Optimized BERT Pretraining Approach', URL: https://arxiv.org/abs/1907.11692.

Morreall J. "Philosophy of Humor", The Stanford Encyclopedia of Philosophy (Fall 2020 Edition). Edward N. Zalta (ed.), Metaphysics Research Lab, Stanford University, 2020, vol. 2. URL: https://plato.stanford.edu/archives/fall2020/entries/humor/

Pritchett Bradley L. Garden Path Phenomena and the Grammatical Basis of Language Processing // Language 64, 1988, p. 539–576.

Raskin V. Semantic Mechanisms of Humor, Volume 24 Springer Netherlands, Dordrecht, 1984, p. 99–147.

Raskin V., Attardo S. Script theory revis(it)ed: joke similarity and joke representation model // Humor – International Journal of Humor Research, Voume. 4 (Issue 3-4), 2020, p. 293–348.

SemEval2020, URL: https://alt.qcri.org/semeval2020/

SemEval2021, URL: https://semeval.github.io/SemEval2021/

Spacy-model "en_core_web_trf": https://huggingface.co/spacy/en_core_web_trf

Tang L., Cai A., Li S., Wang J. The Naughtyformer: A Transformer Understands Offensive Humor, 2023, URL: https://arxiv.org/abs/2211.14369

Toplyn J. Witscript 3: A hybrid ai system for improvising jokes in a conversation. 2023, URL: https://arxiv.org/abs/2301.02695

Veale T. Figure-Ground Reversal in Linguistic Humour:A multimodal prespective // Lodz Papers in Pragmatics 4.1, Special Issue on Humour, 2008, p. 63–81.

Wang M.,Yang H., Qin Y., Sun S., Deng Y. Unified Humor Detection Based on Sentence-pair Augmentation and Transfer Learning // In Proceedings of the 22nd Annual Conference of the European Association for Machine Translation, 2020, p. 53–59.

Weller O., Seppi K. Humor Detection: A Transformer Gets the Last Laugh // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, p. 3621–3625.

Информация об авторе

Заковоротная Евгения Максимовна, аспирант факультета гуманитарных наук Национального исследовательского университета «Высшая школа экономики»

Information about the Author

Eugeniia M. Zakovorotnaia, Postgraduate Student of the Faculty of Humanities of the National Research University Higher School of Economics, Moscow, Russian Federation

Статья поступила в редакцию 14.08.2023;
одобрена после рецензирования 28.03.2024; принята к публикации 19.04.2024

The article was submitted 14.08.2023;
approved after reviewing 28.03.2024; accepted for publication 19.04.2024