ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное автономное образовательное учреждение высшего образования

Национальный исследовательский университет «Высшая школа экономики»

Факультет гуманитарных наук

Образовательная программа «Фундаментальная и компьютерная лингвистика»

Анисимов Арсений Денисович

А ГДЕ СМЕЯТЬСЯ: ДЕЙСТВИТЕЛЬНО ЛИ НЕЙРОСЕТИ СПРАВЛЯЮТСЯ С ЗАДАЧЕЙ HUMOR DETECTION

Выпускная квалификационная работа студента 4 курса бакалавриата группы БКЛ202

Академический руководитель образовательной программы

канд. филологических наук, доц.

Ю.А. Ландер

_____

« » _____ 2024 г.

Научный руководитель

Кандидат филологических наук, доцент.

Д. А. Рыжова

Москва 2024

Аннотация

Юмор, как и многое в психике человека, трудноформализируемая категория. Многие задачи, которые не получается решить с помощью чётких алгоритмов, пробуют решать с помощью глубокого обучения нейросетевых моделей. Обнаружение юмора (humor detection) — одна из таких задач. Главная трудность при обучении моделей для бинарной классификации текстов — противопоставить юмористические тексты в датасете похожими на них неюмористическими. Мы создали датасет A_HDE (Advanced Humor Detection Examples) с текстами на русском языке для обучения и тестирования алгоритмов обнаружения юмора. Датасет состоит из 1608 юмористических текстов и их минимально изменённых 1590 неюмористических версий. На основе современных работ в психологии и лингвистике мы ограничили круг юмористических механизмов, тексты с которыми мы включали в датасет. Модели RuBERT и Conversational RuBERT, обученные на A_HDE и FUN [Blinov et al. 2019], показали низкие результаты на тестовой части A_HDE (Matthew's Correlation Coefficient < 0,199). Мы обнаружили, что модель Conversational RuBERT распознавала один из юмористических механизмом — Garden Path [Dynel 2012] — лучше, чем остальные. Датасет не подходит для полноценного обучения, но может быть использован как бенчмарк для humor detection моделей или для исследований интерпретируемости нейросетей.

1. Введение

Юмор и то, что люди находят смешным, — огромное поле для исследований с позиции разных наук. В сфере обработки естественного языка распознавание юмора (humor detection) является задачей, которая ставит высокие требования как к подбору материала для обучения, так и к способностям модели формализировать концепты. При подборе данных для обучения исследователи обычно ориентируются на количество, а не качество данных. Однако, на наш взгляд, отказ от формализации и от определения того, что понимается под юмором, приводит к тому, что модели учатся видеть поверхностные черты вроде пунктуации, некоторых синтаксических конструкций и эмоционально окрашенной лексики. Мы считаем, что задача нахождения и объяснения юмористических механизмов является важным этапом перед более сложной задачей — генерацией юмора.

Мы предлагаем использовать для обучения моделей инварианты: короткий юмористический текст (1) и несколько текстов с минимальными изменениями, которые превращают изначальный текст в неюмористический (1а) и (1б).

(1) *Электрик, работающий в очень плохих условиях, сказал: «Как-нибудь протяну».*
(1а) *Менеджер, работающий в очень плохих условиях, сказал: «Как-нибудь протяну».*
(1б) *Электрик, работающий в очень плохих условиях, сказал: «Как-нибудь проживу».*

Цели нашей работы — 1) создать датасет на русском языке, состоящий из коротких юмористических и неюмористических текстов. Датасет может быть использован как бенчмарк для моделей, обученных крайне точно идентифицировать юмористические механизмы по минимальным различиям между парами или тройками текстов. 2) Проверить эффективность моделей RuBERT и Conversational RuBERT, обученных на нашем датасете и других датасетах.

Ключевой задачей для нас является ограничить область исследования и подробно описать юмористические механизмы, которые мы собираемся искать. Существует много типов юмора, в том числе анти-юмор и юмор абсурда, обладающие своими уникальными и ещё более трудноформализуемыми чертами.

2. Обзор литературы

2.1. Humor detection: история и проблемы

Несмотря на то, что чувство юмора считается чертой, присущей человеку, компьютерные исследования по обнаружению и генерации юмористических текстов проводились, по крайней мере, с начала 21-го века. В одной из первых работ по обнаружению юмора были применены эвристические алгоритмы машинного обучения [Mihalcea and Strapparava 2005]. В последние десять лет различные архитектуры нейронных сетей стали использоваться для исследования юмора [Ren et al. 2023]. Успешность любой работы по глубокому обучению зависит от количества и качества собранных данных для обучения и тестирования. Качественный набор данных также повышает внешнюю валидность экспериментов.

Авторы недавнего исследования [Ren et al. 2023] обозревают главные работы по изучению юмора в сфере обработки естественного языка. Они анализируют методы, результаты и выводы 29 статей, опубликованных с 2012-го по 2023-й года. Большая часть использованных датасетов содержала только тексты на английском языке, некоторые исследователи проводили эксперименты на данных китайского и немецкого. Датасеты различались по публичной доступности, однако большая часть датасетов была интернет-текстами, собранными из открытых источников, в редких случаях датасет был мультимодальным. Датасеты размечались либо по бинарной классификации

«юмористическое» – «неюмористическое», либо по шкале юмористичности. Датасеты второго типа обычно требовали ручной разметки от нескольких человек, как, например, датасет с соревнований IberLEF 2021 “Humor Analysis based on Human Annotation (HAHA)” (‘Анализ юмора на основе оценки людей’) (Chiruzzo et al., 2021).

В датасетах двух типов разметки юмористические тексты обычно были представлены короткими жанрами: каламбуры (puns), короткие анекдоты, короткие шутки (one-liners). В случае с бинарной разметкой возникает важный вопрос: откуда брать неюмористические примеры. Обычно их также берут из открытых источников, такими текстами могут быть новостные заголовки, пословицы или предложения из национальных корпусов [Ren et al. 2023]. Этот метод, несомненно, эффективен, если цель эксперимента обучить модель различать тексты юмористических жанров среди неюмористических, не исследуя отдельно юмористические механизмы. Модели BERT показали высокие результаты в распознавании юмора на больших датасетах, собранных таким образом. Например, в [Weller and Seppi 2019] дообученный BERT показал F1-score = 0,986 на датасете из 231 657 коротких шуток и 231 657 новостей на английском языке.

В некоторых исследованиях использовались более строгие правила создания датасета, когда исследователи пытались сделать разницу между юмористическими и неюмористическими как можно меньше. Авторы [Peyrard et al. 2021] собрали данные с помощью сайта unfun.me, пользователи которого редактировали заголовки сатирических новостей, чтобы превратить их в реальные новостные заголовки. GPT2 и BERT показали неплохой результат в случае, если модели должны были выбрать из пары заголовков тот, который является сатирическим (метрика accuracy (точности) в зависимости от подкатегории заголовков принимала значения от 0,702 до 0,918 у GPT2 и от 0,723 до 0,989 у BERT). Но вот результаты при предсказаниях лейбла отдельных предложений были гораздо хуже: от 0,493 до 0,553 у GPT2, от 0,582 до 0,704 у BERT. Авторы также выяснили, что одна из голов внимания на слое 10 у BERT была ключевым компонентом в обнаружении сатиры.

Минимальные пары для обнаружения юмора также использовали авторы [Winters and Delobelle 2020]. Они собрали корпус из 3235 шуток на голландском языке и использовали алгоритм dynamic template, чтобы создать неюмористические примеры из юмористических. Алгоритм удалял часть слов из контекста и заменял их другими словами тех же частей речи, отдавая предпочтения менее частотным словам. Авторы отмечают, что в результате юмористические тексты превращались в бессмысленные тексты. Далее исследователи сравнили результаты алгоритма машинного обучения Naïve Bayes, нейросетей архитектур Long Short-Term Memory (LSTM), Convolutional Neural Network

(CNN) и RobBERT на нескольких датасетах: шутки vs. пословицы, шутки vs. новости, шутки vs. dynamic templates. Нейросети показали хорошие результаты на первых двух датасетах: accuracy and F1 > 0,936 у LSTM и CNN, > 0,98 у RobBERT). Однако результаты ухудшились на последнем датасете: ~0,5 у LSTM и CNN; 0,892 у RobBERT).

Пробинг нейросетей показывает, что аккуратный выбор неюмористических примеров важен, иначе нейросети будут опираться на побочные характеристики текстов, такие как: пунктуация, вопросительные слова, именованные сущности [Inácio et al 2023]. Инасио и другие авторы работы «Что выучивают классификаторы юмора?» называют конкретность (concreteness) и вообразимость (imageability) свойствами, которые коррелируют с юмористичностью, однако, это не первостепенные факторы, на которые опираются психологи и лингвисты, изучающие юмор (упоминание фантазии как компонента, помогающего разрешить юмористическое несоответствие есть в [McGhee 1972]).

Определить, что мы имеем виду говоря о словесном юморе и его механизмах, является ключевой задачей нашего исследования. Далее мы кратко перескажем историю изучения словесного юмора и попытки формализации юмористических механизмов.

2.2. История изучения словесного юмора

В этом разделе пойдёт речь об истории изучения юмора, который оперирует в первую очередь устной и письменной формами естественного языка. Безусловно, понятие юмора не всегда было ограничено такими рамками. Изначально латинское *umor* обозначало жидкость в организме. Позже оно стало ассоциироваться с темпераментом человека: холерик, сангвиник, меланхолик, флегматик. В 17–19 веке humor в английском языке обозначало экстравагантную комичность человека. В начале этого пути оно имело пренебрежительный оттенок, а позже утратило его. [Larkin-Galiñanes 2017: 4] Современный смысл слова «юмор» в русском и английском языках — всё, что может вызвать улыбку или смех — слово обрело в 20 веке.

Научное изучение разных аспектов юмора началось не раньше 19 века. До этого «комичное» и природу смеха описывали философы. Самые ранние тексты, дошедшие до нас, принадлежат Платону (5–4 вв. до н. э.) и Аристотелю (4 в. до н. э.). [Larkin-Galiñanes 2017: 5] Они ассоциировали юмор с комедией, считавшейся «низким» жанром. Платон считал, что в сущности смеха лежит злорадство или ощущение превосходства, а смех обычно вызван неудачами «слабых» либо злосчастиями «сильных». Таких же взглядов придерживались ранние христианские авторы (Отцы Церкви, 3–5 вв. н. э.), они порицали юмор и считали смех проявлением пренебрежения. [Larkin-Galiñanes 2017: 6] Взгляд на

юмор как на инструмент высмеивания и унижения получил название «теория превосходства» (Superiority Theory).

Аристотель в «Поэтике» высказал более мягкое мнение: «Смешное — это некоторая ошибка и безобразие, никому не причиняющее вреда и ни для кого не пагубное» [Аристотель: 153]. Аристотель и другие классические авторы разграничивали юмор как высмеивание и как игру слов, придающую выразительность речи. Авторы отмечали, что ораторские приёмы «остроумия» основываются на неожиданности и противоречии. Буквальное прочтение метафорического выражения было языковым приёмом, который философы чаще всего упоминали при классификации разных типов юмора [Attardo 1994: 21].

Теорию несоответствия (Incongruity theory) позже развили в своих работах европейские философы. Иммануил Кант (1724–1804) описывал обнаружение несоответствия в первую очередь как когнитивный процесс, а не эмоциональный. Артур Шопенгауэр (1818–1883) подчёркивает, что важно не просто несоответствие, а нарушение созданного ожидания. Философ Фрэнсис Хатченсон (1694–1746) приводит примеры ситуаций, которые содержат в себе несоответствия нашим представлениям о желаемом мироустройстве: «Есть много несообразностей, которые могут вызвать что угодно, только не смех. Дряхлый человек с тяжелой ношей, пять буханок хлеба и две рыбы среди множества других, и всякая непригодность и грубая диспропорция; расстроенный инструмент, ложка дегтя в бочке меда, майский снег, Архимед, изучающий геометрию во время осады, и все противоречивые вещи; волк в овечьей шкуре, брешь о сделках и обман». [Westbury et al. 2016: 142] Контрпримеры Хатченсона подтверждают, во-первых, мысль Аристотеля, что юмор — «безобразие, не причиняющее вреда», во-вторых, что важной частью для создания юмористического эффекта является наличие ожиданий и разрешение несоответствия. Теория несоответствия получила развитие в психологических и лингвистических работах.

Далее мы сузим область нашего интереса и поговорим о юмористических текстах со структурой сет-ап + панчлайн, она подробно описана далее. Тексты могут быть представлены разными речевыми жанрами: анекдотами, байками, прибаутками [Шмелева, Шмелев 2002], но поскольку в нашем случае это разграничение неважно, мы обобщённо используем название «шутки» для юмористических текстов.

2.3. Психологические теории: что вызывает юмористический эффект в шутках

Джерри Салс утверждает, что для того, чтобы слушатель или читатель оценил шутку, то есть понял её суть, должен произойти двухэтапный процесс обработки

информации. «На первом этапе воспринимающий обнаруживает, что его ожидания относительно текста не подтверждаются концовкой шутки. Воспринимающий сталкивается с несоответствием — этот момент в юмористическом тексте называют кульминацией или панчлайном (the punch line). На втором этапе воспринимающий решает своего рода задачу — ищет когнитивное правило, по которому сказанное в панчлайне следует из основной части шутки и «примиряет» несочетаемые части. Когнитивное правило определяется как логическая пропозиция, определение или факт из опыта. Поиск такой информации позволяет сопоставить несовместимые части шутки. Хотя **эти когнитивные правила трудно свести в систему**, они, по-видимому, являются неотъемлемой частью когнитивного аппарата» [Suls 1972: 82]. Салс подчёркивает, что успешность шутки — то, покажется ли она воспринимающему смешной — зависит от успешного прохождения двух этапов. Разрешение соответствия невозможно без его обнаружения, а только лишь обнаружение не делает шутку смешной[1].

Джерри Салс разбирает прохождение двух этапов на примере, приводим вариацию этой шутки (2).

(2)

Маленькая Этель села за стол и заказала целый фруктовый торт.
— Мне разрезать торт на четыре или восемь частей? — спросила официантка.
— Четыре, — сказала Этель, — мне нельзя много сладкого.

Панчлайном в (2) является часть *мне нельзя много сладкого*. Согласно максиме релевантности [Grice 1975: 45 завершающая фраза Этель должна объяснять, почему Этель попросила разрезать торт на четыре куска. Инконгруэнтность состоит в том, что завершающая фраза противоречит действиям Этель и не объясняет, почему она попросила разрезать торт так. Она не понимает или игнорирует, что куски будут различаться в размере, а размер всего торта не изменится. Вместо этого Этель использует «эвристическое правило, согласно которому увеличение количества приводит к

[1] Мы должны оговорить, что возможность разрешения несоответствия не является обязательной для (а) всех воспринимающих, (б) всех типов юмора. Психолог Пол МакГи отмечает, что дети до 7–8 лет не обладают достаточными когнитивными способностями, чтобы проходить второй этап и полностью понимать шутки, но находят забавной саму инконгруэнтность. Например, дети дошкольного возраста могут смеяться над «туалетным юмором», где инконгруэнтностью является нарушение табу на озвучивание темы. [McGhee 1972: 77] Также юмор абсурда и некоторые другие типы юмора могут не предполагать разрешение инконгруэнтности/

увеличению общего объема» [Suls 1972: 83]. Правило можно представить в виде цепочки рассуждений:

‘больше кусков торта (восемь)’ => ‘больший объём торта’ => ‘плохо’.

‘меньше кусков торта (четыре)’ => ‘меньший объём торта’ => ‘хорошо’.

Разрешение несоответствия происходит, когда слушатель / читатель восстанавливает цепочку рассуждений и осознаёт, что Этель использует логичное правило, но намеренно или ненамеренно применяет его в неподходящей ситуации.

Изменяя финальную часть шутки, Салс показывает, что «несоответствие — необходимое, но недостаточное условие» для успешного достижения юмористической цели текста. Такие ответы Этель как (2а) и (2б) были бы инконгруэнтными, но несмешными, так как они не дают прозрачной возможности для выведения когнитивного правила, связывающего сет-ап и панчлайн.

(2а) — *Не режьте. Мне нельзя много сладкого*.
(2б) — *Восемь кусочков. Мне нельзя много сладкого*.
[Suls 1972]

Если бы Этель неиллюзорно придерживалась озвученного правила, то шутка теряла бы основу для возникновения несоответствия. Этель могла бы дать конгруэнтный несмешной ответ, например (2в).

(2в) *На четыре. Мне нельзя много сладкого, а торт для мамы*.

Особенно важно отметить, что юмористические тексты остаются инконгруэнтными после разрешения центрального несоответствия, так как когнитивное правило само привносит новую инконгруэнтность. «Обычно, если инконгруэнтность полностью разрешается, то юмористический эффект не возникает» [Forabosco 2008]. Авнер Зив подчёркивает, что локальная логика шутки (local logic, аналог «когнитивного правила» Салса) позволяет замаскировать абсурдность, которая вызывает юмористический эффект [Ziv 1984]. В шутке (2), как было показано ранее, несоответствие проявляется в незамаскированной абсурдности рассуждений Этель, использующей принцип ‘меньше кусков => меньший объём’.

Нецентральная инконгруэнтность — то есть та, которая не разрешается на втором этапе в модели Салса — может выводиться из неявных компонентов значения. Пример — (3).

(3) *Outside of a dog, a book is man's best friend. Inside of a dog, it's too dark to read.*

Groucho Marx (1890–1977) [Aarons 2017: 84]

(Игра слов непереводима на русский язык. См. объяснение ниже)

Шутка основана на многозначности выражения *outside of a dog*. Оно может быть воспринято в переносном значении 'не считая собаки' и в прямом значении 'снаружи, вне собаки'. Более вероятной интерпретацией первого предложения будет 'не считая собаки'. Второе предложение заставляет воспринимающего переключиться на более буквальную интерпретацию за счёт параллелизма слов *outside – inside*, которые являются антонимами в прямых значениях. Инконгруэнтность привносится за счёт абсурдности идеи, что кто-то пытается заглянуть внутрь собаки. «Дополнительная инконгруэнтность возникает из-за объяснения, почему внутри собаки невозможно читать — там слишком темно. То есть предполагается, что, если бы она была освещена, проблем бы не возникло». [Aarons 2017: 85] То есть восприятие такого рассуждения и нахождение в нём смысла заставляет слушателя / читателя вовлекаться в «игру» по правилам локальной логики.

Таким образом, сформулируем главные определения для нашего исследования.

Инконгруэнтность (несоответствие) — ощущение несогласованности между двумя частями юмористического текста. Инконгруэнтность представляет собой некоторую логическую задачу, решение которой позволит слушателю / читателю увидеть согласованность между частями.

Разрешение центральной инконгруэнтности (несоответствия) — успешное решение логической задачи. Логическая задача представляет собой нахождение некоторой неявной, локальной логики [Ziv 1984] и восстановления рассуждений, основываясь на такой логике. В некоторых случаях локальная логика предполагает ошибочный, абсурдный или непрагматичный ход мыслей или действий.

Панчлайн или кульминация — часть юмористического текста, осознание которой позволяет слушателю / читателю разрешить центральную инконгруэнтность в тексте. Панчлайн находится в конце текста, границы сет-апа и панчлайна определяются с помощью разделения текста на две части. Финальная часть считается панчлайном, если при её отрезании первая часть оказывается несмешной, то есть, по модели Салса, не разрешает центральное несоответствие [Hockett's (1973); Aljared 2017: 67].

Сет-ап — часть юмористического текста до панчлайна.

2.4. Лингвистические теории: как может быть построен юмористический текст

Многие исследователи юмора, прежде всего лингвисты, делят шутки на вербальные и невербальные. Краткий обзор принципов разделения приводится в [Aljared 2017: 67]. Как и Альма Альджаред, мы руководствуемся определениями Аттардо [Attardo 1994: 95] для вербального и референциального (невербального) юмора. Аттардо приводит критерии, по которым это разделение выполняется.

Шутки референциального типа «основаны исключительно на смысле текста и не содержат никаких ссылок на фонологическую реализацию лексических единиц». Вербальные шутки содержат как ссылку на фонологическую реализацию какого-либо элемента, так и на смысл текста [Аттардо 1994: 95]. Арво Крикманн [Krikmann 2006] отмечает, что из этого можно сделать вывод, что референциальные шутки легко переводимы на разные языки, а вербальные шутки разве что могут оказаться переводимыми по совпадению. В качестве примеров референциальной и вербальной шутки служат (2) и (3) соответственно.

Однако многие лингвистические теории юмора имеют свои белые пятна. Марта Дайнел показывает, что главный недостаток многих теорий, разработанных в конце прошлого и в начале этого века, состоит в их неприменимости к шуткам с разными юмористическими механизмами [Dynel 2012: 25].

Марта Дайнел выделяет три юмористических механизма шуток: garden path, red light и crossroad. Они различаются тем, когда в тексте оказывается представлена центральная инконгруэнтность и когда она разрешается. [Dynel 2012]

Garden Path («садовая дорожка», GP). Сет-ап заставляет слушателя / читателя прийти к легкодоступному толкованию ситуации, которое оказывается неверным после восприятия панчлайна. Панчлайн раскрывает скрытое толкование сет-апа, выявляя не замеченную ранее двусмысленность[2]. В зависимости от длины шутки двусмысленным выражением может быть либо весь сет-ап, либо его часть. Марта Дайнел отмечает, что структура шуток с эффектом садовой дорожки описывается чаще всего в теориях словесного юмора [Dynel 2012: 25].

---

[2] Неявность двусмысленности — то, что отделяет «хорошую» шутку от «плохой». Под плохой шуткой обычно понимается предсказуемая. Мы не ставим задачу оценить степень предсказуемости шуток и их качество.

(4)

— *Ты чего такой хмурый?*

— *Да сегодня один раздолбай меня зуба лишил и деньги отобрал!*

— *Гопник что ли?*

— *Не, стоматолог!*

S1 ‘гопник выбил зуб в драке и отобрал деньги’ & S2 ‘стоматолог не смог спасти зуб пациента и вырвал его, но пациенту пришлось заплатить’ = Ø

Дизъюнктор = *Не, стоматолог.*

Коннектор = *один раздолбай меня зуба лишил и деньги отобрал.*

Важной частью шуток типа GP являются элементы, которые делают возможным переключение с одной интерпретации на другую, и наложение двух интерпретаций. В работе [Attardo 1994] эти элементы описываются с помощью isotopy disjunction model (букв. «модель разъединения изотопии», IDM). Сальваторре Аттардо определяет изотопию как смыслы текста [Aljared 2017; Attardo 1994: 96]. В случае с шутками типа GP слушатель / читатель воспринимает текст линейно и выводит некоторый смысл текста S1. Затем слушатель / читатель встречается с неожиданным элементом — **дизъюнктором**, противоречащим смыслу S1 и заставляющим переключиться на другой смысл, S2. Смыслы S1 и S2 в каком-то аспекте противоречат друг другу. Слушатель / читатель мысленно возвращается к предыдущему тексту и находит коннектор. **Коннектор** — некоторое выражение, позволяющее совместить текст как со смыслом S2, так и со смыслом S1. Аттардо использует термин «коннектор» только применительно к вербальному юмору, однако та формулировка, в которой он определяет коннектор, на наш взгляд, применима по отношению к референциальному юмору и облегчает его классификацию. Разберём функционирование коннектора на примере референциальной шутки (5).

(5) *Я понял, что Деда Мороза не существует, когда жена попросила подложить подарки сыну под ёлку.*

S1 ‘персонаж говорит о времени, когда был ребёнком’ & S2 ‘персонаж говорит о времени, когда он взрослый’ = Ø

Коннектор = *Я понял, что Деда Мороза не существует*(*, когда*)

Дизъюнктор = *когда жена* (*попросила подложить подарки сыну под ёлку*).

Основываясь на жизненном опыте, слушатель / читатель воспринимает фразу-коннектор *Я понял, что Деда Мороза не существует*(*, когда*) с импликатурой S1, но позже вынужден принять толкование S2. В этой шутке коннектор связывает две интерпретации, но не является лексически или синтаксически многозначным выражением.

В шутках типа garden path две интерпретации обязательно в чём-то противоречат друг другу. Шутка с механизмом «садовой дорожки» работает только тогда, когда смысл S1 более явный и контекст подкрепляет его активацию, а смысл S2 наоборот, трудно вычленим. Это возможно за счёт неограниченного набора прагматических средств. В рассмотренных ранее примерах это были: отсылка к известным выражениям (‘*Собака — друг человека*’ в (3)), выбор лексики и синтаксических конструкций (4). Шутки с GP-механизмом часто используют не только прагматическую неоднозначность, но и полисемию, омонимию и деидиоматизацию некомпозициональных выражений [Dynel 2012: 11]. Деидиоматизация может возникать как на уровне одного слова или словосочетания (6), так и на уровне нескольких расположенных рядом слов (7).

(6) *В городе открыто новое месторождение — роддом.*

S1 ‘шахта или карьер’ & S2 ‘роддом’ = Ø

Дизъюнктор = *роддом.*

Коннектор = *месторождение* (неявн. *место рождения*)

(7) Штирлиц играл в карты и проигрался. Но Штирлиц умел делать хорошую мину при плохой игре. Когда Штирлиц покинул компанию, мина сработала.

S1 ‘скрывать разочарование & S2 ‘установить взрывное устройство’ = Ø

Найденные нами анекдоты о Штирлице, в основном, относятся к типу garden path. А. С. Архипова отмечает, что структура каламбурных анекдотов о Штирлице непротиворечиво описывается с помощью формальной теорий юмора Script-Based Semantic Theory of Humor [Raskin 1985] ([Dynel 2012: 23–25] отмечает сходство SSTH Раскина и IDM Аттардо). Анекдоты о Штирлице часто строятся на абсурдной ре-интерпретации устойчивых выражений как имён собственных. [Архипова 2003: табл. 4]

(8) *Штирлиц стрелял вслепую. Слепая упала назвзничь. Взничь бросился наутек. Утек бежал первым.*

Red Light («красный свет», RL). Название «красный свет» построено на следующей метафоре: слушатель / читатель воспринимает определённую интерпретацию, пока ему не приходится неожиданно остановится на красном свете (что соответствует кульминационному моменту). Панчлайн открывает возможность увидеть вторую, скрытую интерпретацию. Однако после кратковременной паузы слушателю / читателю не приходится отказываться от первоначальной интерпретации. «Пауза часто остается незамеченной, если устранение несоответствия не создает серьезных проблем». Панчлайн вводит центральную инконгруэнтность и сразу же её разрешает. [Dynel 2012: 16] Разберём работу юмористического механизма RL на примерах.

(9)

— Спасите!

— А?

— Если вы меня сейчас не вытащите, я утону!

— Не, ну ты смотри, он ещё условия ставит!

Дизъюнктор — последняя реплика, коннектор — предпоследняя.

В (9) слушатель / читатель наблюдает за развитием ситуации. Когда он доходит до панчлайна — последней фразы в диалоге — он не вынужден переосмыслять интерпретацию прочитанного ранее. Он по-прежнему представляет диалог двух людей, один из которых тонет и пытается сделать так, чтобы его спасли. Инконгруэнтность состоит в том, что второй говорящий воспринимает поверхностную форму того, что говорит первый, как ультиматум, а не крик о помощи. Возвращаясь к предпоследней реплике в диалоге, слушатель / читатель убеждается в том, что предложение действительно сформулировано как условие, однако оно всё ещё остаётся криком о помощи — интерпретация сет-апа не меняется. Пользуясь определениями Аттардо [Attardo 1994: 96], в текстах такого типа мы также может выделить коннектор, связывающий две интерпретации, и дизъюнктор — то, что даёт возможность воспринимающему вычленить неявную интерпретацию.

Шутки с механизмом red light дополнительно отличаются от шуток типа garden Path тем, что коннектор в RL может быть расположен как до дизъюнктора, так и после него. В шутках garden path дизъюнктор всегда следует за коннектором. Пример шуток типа RL с расположением коннектор —> дизъюнктор (9), с расположением дизъюнктор —> коннектор (10).

(10)

— *Будете вино?*

— *Нет.*

— *Я настаиваю.*

S1/2 ‘настоятельно прошу’ & S1/2 ‘изготавливаю вино’

Коннектор — *Я настаиваю*.

В случаях с появлением коннектора в панчлайне возможно только предсказать, а не предопределить, какая из интерпретаций у слушателя / читателя возникнет первой (прим. (10)). Иногда в вербальных шутках типа RL одна из интерпретаций не является полноценным, существующим в языке смыслом как в шутках GP. Шутка может быть основана на созвучии ассоциативно схожих частей, быть игрой слов (прим. (11)).

(11) *Отмычка — приспособление, имеющее для вора ключевое значение.*

В (11) нет явного дизъюнктора, а прилагательное *ключевое* является коннектором интерпретаций ‘первостепенный’ и ‘относящийся к ключу’.

Crossroad («распутье», C). Единственный из юмористических механизмов, в котором инконгруэнтность вводится до панчлайна. «Распутье» — метафора того, что слушатель / читатель не может выбрать из необозримого количества альтернатив, по какому из путей интерпретации пойти, пока не разрешит инконгруэнтность с помощью информации в панчалайне. [Dynel 2012: 13] Шутки типа C часто формулируются в форме псевдозагадок (12), но также могут быть представлены в виде более длинных историй или диалогов (13).

(12) *Почему не стоит бить кассиршу? Она может дать сдачи.*

(13)

Бард приезжает домой после концерта с огромным фингалом под глазом. Жена спрашивает:

— *Что произошло?*

— *Цветы на сцену кидали…*

— *Ну и что?!*

— *Один из них летел горшком вперед…*

Коннектор — *Цветы на сцену кидали*.

В (12) коннектор, который является многозначным выражением, расположен в панчлайне — ответе на загадку. Явно выраженного дизъюнктора, как и в случае с (11), в шутках такого типа нет.

Анекдот (13) устроен чуть более сложно. Слушатель / читатель до того, как услышит / прочитает последнюю реплику, сталкивается с задачей — фингал под глазом не объясняется тем, что на сцену кидали цветы. В отличие от шуток типа garden path слушателю / читателю не доступно толкование, которое легко обнаружить. Однако фраза *цветы на сцену кидали* имеет ключевое значение для того, чтобы разрешить инконгруэнтность с помощью финальной реплики. Как и в случае с шутками GP фраза раскрывает свою многозначность после панчлайна, поэтому мы используем термин «коннектор» в том числе для описания шуток такого типа, несмотря на то, что начальная теория их не описывала [Dynel 2012: 25].

Если шутка типа Crossroad сформулирована в одном предложении, то часто она представлена в виде текста, в котором тяжело этом полностью исключить инконгруэнтность, заменив малое количество слов, прим. (14).
(14) *Из-за шизофрении распалась труппа театра одного актёра.*

Описания юмористических механизмов garden path, red light и crossroad, а также роли коннектора и дизъюнктора важно для нас, чтобы обозначить границы исследуемой области и определить критерии для отбора юмористических текстов в датасет.

3. Методология

3.1. Составление датасета A_HDE

3.1.1. Принципы отбора

Датасет A_HDE (Advanced Humor Detection Examples) состоит из коротких юмористических текстов и их изменённых неюмористических версий — «минимальных пар». В датасете 1608 текстов с лейблом 1 (шутки) и 1590 текстов с лейблом 0 (не-шутки).

Мы брали шутки из открытых источников: сайтов anekdot.me, anekdot.ru, anekdoty.ru, myanekdot.ru, ruanekdot.ru, vse-shutochki.ru, публикации на dzen.ru и группа «Сборник тупых каламбуров» vk.com/puns_about_sewing. Шутки должны были содержать коннектор — слова или выражения, позволяющее воспринять текст в двух разных интерпретациях (о термине «коннектор» см. раздел 2.4). Также обе интерпретации должны

были быть выводимы за счёт порядка преподнесения информации или за счёт других слов, то есть контекста. Существуют работы, в которых похожие принципы формализированы. Например, в [Kao et al. 2015] предлагается формальная модель для фонетических каламбуров (phonetic puns), то есть шутки, содержащие «слова, которые звучат одинаково с другими словами (английского языка) или похоже на них». Для модели важны два параметра каламбура: двусмысленность (ambiguity) и различимость (distinctiveness). Двусмысленность означает, что контекст допускает употребление любого из омофонов. Различимость означает, что в контексте есть слова, активизирующие каждое из значений отдельно. Авторы приводят формулы, для расчёта каждого параметра [Kao et al. 2015: 1274]. Отличие нашей методологии состоит в том, что мы брали в датасет шутки, где выведении одной или обеих интерпретаций невозможно без допущения синтаксических нарушений. Е. В. Рахилина называет несовершенство каламбура — «не полное совпадение реально встречающихся в языке форм» — параметром, который включён в правила построения нетривиального каламбура [Рахилина, Плунгян 2009: 145–146]. При разработке датасета мы также отказались от строгой числовой формализации наподобие представленной в [Kao et al. 2015].

Мы не брали в датасет шутки некоторых типов. Далее мы приведём примеры таких шуток на русском и английском языке.

1. Шутки, не имеющие коннектора и не разрешающие инконгруэнтность. Например, сравнения (15). Инконгруэнтность сравнения тяжело измерить и провести границу между поэтической метафорой и комичным сравнением.

(15) *Time flies like an arrow. Fruit flies like a banana.*
(Groucho Marx (1890–1977), из статьи [Aarons 2017])

2. Аллитеративные шутки. То есть юмористические тексты с игрой слов, основанная на далёком сходстве фонетической или графической формы.

(15) *Начали ремонт в стиле "хай-тек", продолжили в стиле "пусть-так".*

3. Шутки, в которых юмористический эффект строится на передразнивании, пародировании произношения, например (16). Неправильное произношение передаётся на письме, что делает текст нечитаемым для моделей, которые не кодируют фонетическую информацию. Если главная инконгруэнтность шутки не строилась на неправильном произношении, то мы редактировали текст и брали его в датасет.

(16)

Заходит парень в ресторан. Заказывает еду и говорит:

Можно мне немного соуса чили?

Официант ему:

— Извините, но это японский ресторан.

*Парень: *растягивает глаза**

— Мозьня мне нимнога соуся цили??

4. Шутки, в которых образовывалось новое слово, например (17). Мы посчитали, что такие шутки были бы явной подсказкой для модели ориентироваться на токены, не присутствующие в словаре её токенизатора.

(17) Q: What do you get when you cross a cow and a lawnmower? A: A lawnmooer.

‘Что получится, если скрестить корову и газонокосилку? Газонокосилка + *му*’.

3.1.2 Концепция минимальных инвариантов и принципы замены юмор –> неюмор.

Минимальная пара в нашем исследовании — противопоставление юмористического и неюмористического текста, где неюмористический текст был получен при минимально возможном изменении неюмористического. Возможная степень минимальности — субъективный фактор, он определялся для каждой шутки отдельно при создании датасета.

При создании датасета мы столкнулись с проблемой быстрого отбора юмористических текстов и изменения некоторых отобранных текстов, поэтому решили придерживаться принципа парности менее строго. Некоторые шутки мы изменяли несколько раз, в следствии чего получили тройки 1–0–0 или четвёрки 1–0–0–0 (0 — неюмор. текст, 1 — юмор.). Некоторые очень похожие шутки (различающиеся максимум двумя леммами; порядок и количество остальных токенов одинаковы) мы также объединяли в четвёрки 1–0–1–0. Далее мы будем называть минимально различающиеся сочетания двух-трёх-четырёх текстов **инвариантами**.

Номер инварианта для каждого текста указан в отдельном столбце. Всего в датасете 1652 инварианта.

Чтобы сбалансировать большее количество неюмористических текстов, мы не удаляли непарные шутки — те, которые не смогли минимально изменить, придерживаясь

алгоритма (см. далее). Мы отметили 118 непарных шуток в отдельном столбце, чтобы их было легко исключить, если эта опция потребуется.

Далее мы приводим алгоритм изменения юмористического текста на неюмористический. Этот же алгоритм опубликован в нашем репозитории https://github.com/arsen-geek/a_hde и может быть использован другими исследователями при создании аналогичного датасета или пополнении нашего.

«Сломать шутку» — создать неюмористический вариант того же текста, заменив в нём как можно меньше токенов. Это можно сделать двумя способами: заменить коннектор (многозначное выражение) (18а) или часть панчлайна (18б).

(18) *Не среда определяет бытие, а пятница.*
(18а) *Не четверг определяет бытие, а пятница.*
(18б) *Не среда определяет бытие, а сознание.*

Для большинства шуток мы создавали по 1–2 сломанных вариантов каждым способом, но оба способа не получится применить везде. Итого сломанных вариантов одной шутки получается от одного до четырёх.

От более предпочтительного варианта изменения к менее предпочтительному:

1) Заменить как можно меньше слов, сохраняя их часть речи. Пример (18а), (18б).

Если 1 не получается, то:

2) Сохранять примерное кол-во слов и пунктуацию (+– 2 токена). Пример (19а).

(19) *Раньше я чувствовал себя мужчиной в женском теле. А потом родился.*
(19а) *Раньше я чувствовал себя мужчиной в женском теле. А потом поменял пол.*

Рекомендуется заменять на слова с тем же корнем, который есть в сет-апе или многозначные по той же теме (*добился* в (20а)).

(20)

После разгромного поражения боксёр приходит в раздевалку — нос сломан, зубы выбиты, сотрясение мозга. Сидит, страдает.

Заходит тренер:

— У меня для тебя прекрасная новость!

— Какая?

— Я договорился на завтра о матче-реванше.

(20а)

После разгромного <…>

— Я добился дисквалификации твоего соперника.

Особое внимание уделять отрицательным слова *нет*, *не*, *нисколько* и т. д. В этих случаях создавать неюмористические варианты с отрицанием и без отрицания. Если разметчик предлагает несмешной вариант, не сохраняя отрицание, то при написании несмешного варианта другой шутки будет необходимым добавить хотя бы один вариант с отрицанием.

(21) *— Вы страдаете алкоголизмом? — Нет, наслаждаюсь.*

(21а) *— Вы страдаете алкоголизмом? — Нет, не пью.*

(21б) *— Вы страдаете алкоголизмом? — Нет, наркоманией.*

Если п. 2 не получается выполнить, то:

3) Заменить на часть текст, подходящую по смыслу, но не сильно длиннее или короче. Пример (22а).

(22)

Приходит Вовочка в первый класс.

— А ты считать умеешь? — интересуется учительница.

— Конечно! Один, два, три, четыре, пять, шесть, семь…

— А дальше знаешь?

— … восемь, девять, десять, валет, дама, король, туз.

(22а)

Приходит Вовочка <…>

— … восемь, девять, десять, одиннадцать, двенадцать, тринадцать, четырнадцать…

Примеры (1) в разделе 2.3 (повторяем их здесь как (23а) и (23б)) показывают, что изменённые тексты шуток могут сохранять инконгруэнтность, но не давать достаточной информации для её разрешения, а соответственно такие тексты не будут юмористическими в модели Дж. Салса.

В датасете мы дополнительно разметили инконгруэтные неюмористические тексты, всего их 41.

(23)

Маленькая Этель села за стол и заказала целый фруктовый торт.

— Мне разрезать торт на четыре или восемь частей? — спросила официантка.

— Четыре, — сказала Этель, — мне нельзя много сладкого.

(23а) *— Не режьте. Мне нельзя много сладкого.*

(23б) *— Восемь кусочков. Мне нельзя много сладкого.*

[Suls 1972]

Мы отмечали, что неюмористический текст содержит инконгруэнтность, руководствуясь следующими критериями:

1. Структура предложения подразумевает причинно-следственную связь между двумя пропозициями, но к наличие этой связи совершенно неочевидно или ложно.

(24) *Электрик решил устроиться в ритуальное агентство, ведь он эксперт по розеткам.*

=> Как связана специализированные знания о розетках и решение работать в ритуальном агентстве?

(25) *Всегда восхищался филологами, у них прекрасная физическая подготовка.*

=> Разве для филологов характерна хорошая физическая подготовка?

2. Речь о реальном мире, и по меркам него ситуация абсурдная.

(26) *Суровые челябинские полицейские задержали дерево. По их словам, у дерева был пистолет.*

(27) *Новости науки. Британские ученые разбили очередной ящик виски.*

=> Почему это относится к «новостям науки»?

3. Недостаточно информации, чтобы объяснить высказанную мысль.

(28) *Еле-еле — наречие, приводящее шахматистов в ступор.*

(29) *Будущих банкиров родителям приносит аист.*

3.1.3. О параметрах разметки на основе лингвистических теорий юмора

Для каждого юмористического текста мы указывали тип юмористического механизма по классификации Марты Дайнел (см. раздел 2.4, [Dynel 2012]): garden path (GP), red light (RL), crossroad (C). Мы отмечали шутки, в которых коннектор находился в сет-апе. По умолчанию мы считали, что коннектор находится в панчлайне.

В датасете присутствуют как вербальные, так и референциальные шутки, эти типы отдельно не размечались. Некоторые вербальные шутки основаны на игре с внутренней формой слова, они отмечены в столбце с комментариями.

Мы разделили текст каждой шутки на сет-ап и панчлайн автоматически. Если текст шутки состоял из нескольких предложений, то последнее предложение выделялось как панчлайн. Если текст шутки представлял собой одной предложение, то панчлайном считалась либо часть от предпоследнего знака препинания до конца текста, либо последние четыре токена (когда единственный знак препинания был в конце текста).

Мы удалили посторонние ссылки и латинские символы внутри слов русского языка. На некоторых сайтах кириллические символы в шутках были заменены на одинаковые по внешнему виду латинские символы, чтобы повысить уникальность текста или зацензурировать обсценную лексику.

3.1.4. Использование датасета FUN

В нескольких экспериментах мы обучали модели на русскоязычном датасете FUN [Blinov et al 2019]. Тренировочная часть FUN состоит из 251 416 текстов (125 708 юмористических и 125 708 неюмористических). Юмористическая часть была собрана с помощью краулинга сайтов VK.com и anekdot.ru (477K), неюмористическая — с помощью краулинга форума Екатеринбурга E1.ru. Шутки и посты с форму сопоставлялись по сходству лексики по алгоритму BM25.

3.2. О моделях

BERT (Bidirectional Encoder Representations from Transformers) — модель-трансформер, предобученная на текстах с частью замаскированных слов для задач предсказания следующего слова и следующего предложения. Модель относится к классу encoder-only [Devlin et al. 2018].

Мы тестируем работу двух моделей: RuBERT и Conversational RuBERT, разработанными лабораторией Deep Pavlov МФТИ [Kuratov, Arkhipov 2019]. Они одинаковы по архитектуре: дообучены на текстах на русском языке с различением строчных и заглавных букв, состоять из 12 слоёв и имеют 12 голов на каждом слое.

RuBERT был дообучен на текстах из русской части Wikipedia и текстах новостей. Conversational RuBERT отличается тем, что его дообучили на текстах субтитров OpenSubtitles, форумах d3.ru и pikabu.ru, подкорпусе текстов из социальных сетей Taiga. Мы предположили, что Conversational RuBERT справится с задачей humor detection на A_HDE лучше, чем RuBERT, потому что шутки по жанру ближе к жанрам текстов, на которых обучался Conversational RuBERT.

Для экспериментов мы использовали вариацию BERT for sequence classification (BERT для классификации последовательности).

3.3. Метрика

Чтобы оценить, насколько хорошо модель распознаёт юмор, мы используем коэффициент корреляции Мэттью (Matthew's correlation coefficient, MCC). Коэффициент принимает значения от –1 до 1. –1 означает, что предсказания модели полностью противоположны правильным ответам. 0 — предсказания модели случайны. 1 — предсказания модели полностью соответствуют правильным ответам.

Коэффициент корреляции Мэттью в отличие от accuracy подходит для тестовых выборок, несбалансированных по количеству элементов в классах.

Мы предпочли использовать MCC вместо F_1-score, так как F_1-score в большей степени зависит от истинно положительных ответов, а для нас особенно важно измерить качество предсказаний на неюмористических примерах. MCC равномерно учитывает все категории ответов при оценке бинарного классификатора.

$$\mathrm{F1} = \frac{2TP}{2TP + FN + FP}$$

$$\mathrm{MCC} = \frac{TN \times TP - FP \times FN}{\sqrt{(TN + FN)(FP + TP)(TN + FP)(FN + TP)}}$$

Формулы метрик F_1-score и Matthew's correlation coefficient. Классы предсказаний: TP – true positive (истинно положительный), FP – false positive (ложно положительный), TN – true negative (истинно положительный), FN – false negative (ложно положительный).

4. Результаты

4.1. Эксперименты с обучением и тестированием на A_HDE

Мы обучали и валидировали RuBERT и Conversational RuBERT на 80% датасета A_HDE (2558 текстов: 1297 шуток, 1261 не-шуток). После каждой эпохи обучения мы тестировали модель на оставшихся 20% A_HDE (640 текстов: 311 шуток и 329 не-шуток).

Авторы рекомендуют обучать BERT 2–4 эпохи [Devlin et al. 2018], однако такое обучение не позволило нам сильно повысить метрику качества. Мы обучали RuBERT 5 и 8 эпох.

RuBERT	1	2	3	4	5	6	7	8
5 эпох	0	0,054	0,067	0,075	**0,103**			
8 эпох	0,02	0,044	0,077	0,066	0,106	0,073	**0,122**	0,096

Табл. 1. Значения Matthew's correlation coefficient модели ruBERT на тестовой выборке A_HDE после каждой эпохи обучения. NB: тестовая выборка использовалась исключительно для измерения качества и не влияла на корректировку весов модели.

Также мы решили обучить RuBERT за 12 эпох, сильно отклонившись от рекомендаций авторов [Devlin et al. 2018]. Метрика MCC повысилась, однако корреляция предсказаний модели и истинных ответов всё ещё очень слабая.

	1	2	3	4	5	6	7	8	9	10	11	12
RuBERT	0,036	0,001	0,071	0,117	0,133	**0,153**	0,111	0,149	0,139	0,125	0,138	0,147
Conv. RuBERT	0,111	0,047	0,125	0,156	0,182	0,178	0,128	0,170	**0,196**	0,179	0,171	0,173

Табл. 2. Значения Matthew's correlation coefficient двух моделей на тестовой выборке A_HDE после каждой эпохи обучения на тренировочной выборке A_HDE.

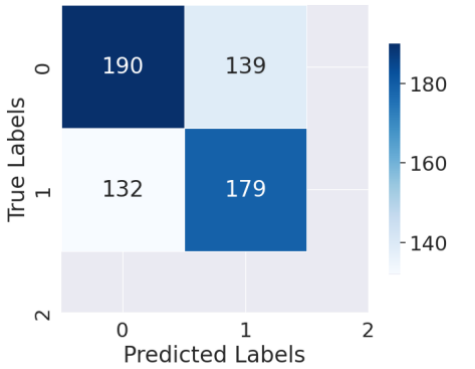

Рис. 1. Соотношение правильных ответов (True Labels) и ответов предсказаний RuBERT (Predicted Labels) после 6-й эпохи обучения (MCC = 0,153).

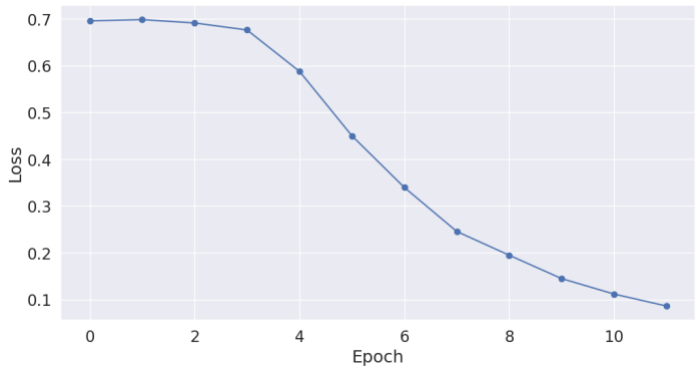

Рис. 2. График функции потерь после каждой эпохи обучения.

Conv Rubert 9 эпоха (MCC=0,196). График функции потерь после на каждой эпохе.

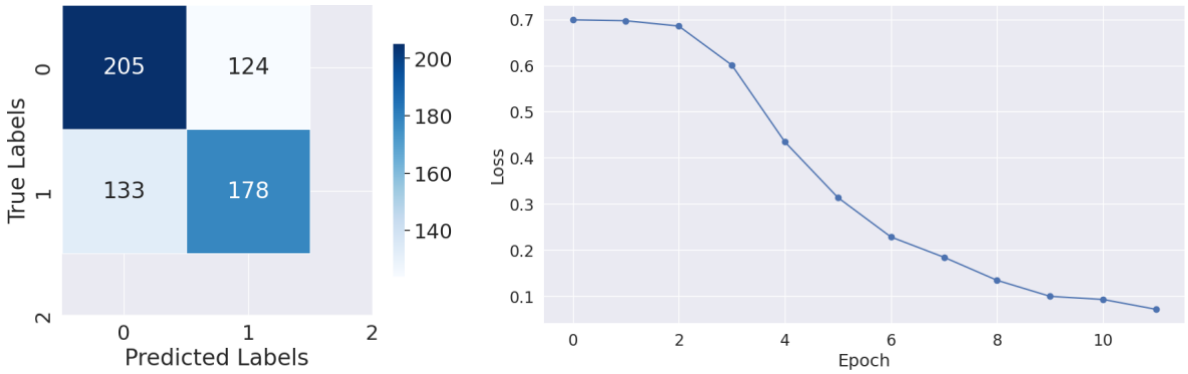

Рис. 3. Соотношение правильных ответов (True Labels) и ответов предсказаний Conversational RuBERT (Predicted Labels) после 9-й эпохи обучения (MCC = 0,196).

Рис. 4. График функции потерь после каждой эпохи обучения.

Мы удалили из обучающей и тестовой выборки неюмористические тексты, в которых сохранена инконгруэнтность (подробнее в разделе 3.1.2), однако это не улучшило предсказания модели.

	1	2	3	4	5	6	7	8	9	10	11	12
Conv. RuBERT	0,035	0,062	0,044	0,132	0,13	0,125	0,141	0,15	0,16	0,145	0,157	**0,169**

Табл. 3. Значения Matthew's correlation coefficient модели Conversatonal RuBERT на тестовой выборке A_HDE после каждой эпохи обучения на тренировочной выборке A_HDE. Из обеих выборок были удалены неюмористические примеры, содержащие инконгруэнтность.

4.2. Эксперименты с обучением на FUN + A_HDE и тестированием на A_HDE

Мы обучили модели на большом датасете FUN и на 80% A_HDE. Количество эпох обучения на FUN варьировалось от 2 до 4, на A_HDE от 1 до 12. Мы измеряли качество предсказаний моделей на 20% A_HDE при всех вариантах обучения.

Обучение RuBERT на 2–4 эпохах FUN + A_HDE не повлияло на перфоманс, если сравнивать с RuBERT, обученном только на A_HDE. Максимальный MCC за все эпохи обучения на FUN и A_HDE достиг всего 0,154.

Обучение Conversational RuBERT на FUN также не помогло повысить качество. В первом эксперименте Conversational RuBERT показал MCC на 0,046 пункта выше, чем RuBERT. В нынешнем эксперименте лучший результат Conversational RuBERT'а MCC = 1,99 отличался примерно так же от результата RuBERT'а MCC = 0,154.

Conv. ruBERT	1	2	3	4	5	6	7	8	9	10	11	12
3 эп. на FUN + 12 эп. на A_HDE	0,05	0,025	0,12	0,15	0,183	0,171	**0,199**	0,173	0,18	0,177	0,185	0,195

Табл. 4. Значения Matthew's correlation coefficient двух моделей Conversatonal RuBERT на тестовой выборке A_HDE после каждой эпохи обучения на тренировочной выборке A_HDE.

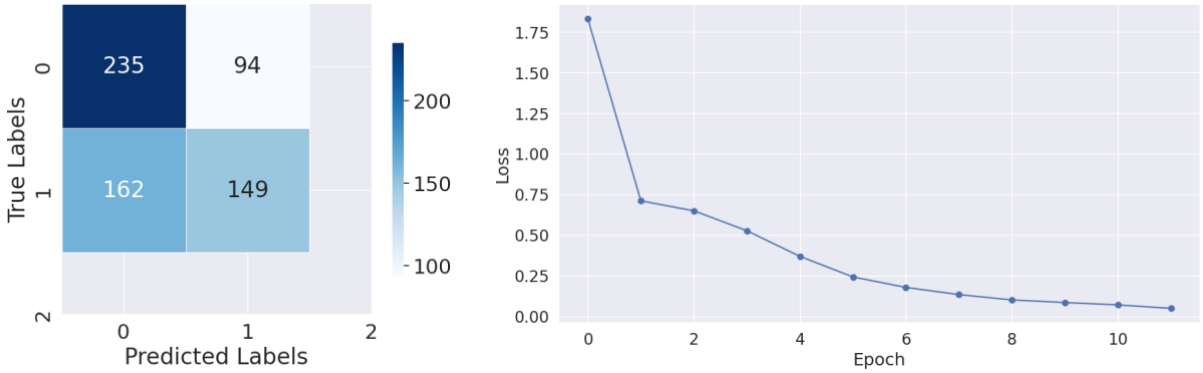

Рис. 5. Соотношение правильных ответов (True Labels) и ответов предсказаний Conversational RuBERT (Predicted Labels) после 7-й эпохи обучения (MCC = 0,199).

Рис. 6. График функции потерь после каждой эпохи обучения на A_HDE.

Так же, как и в эксперименте 4.1, мы удалили неюмористические инконгруэнтных примеры из тренировочной и тестовой выборки. Их исключение не повлияло на результаты.

4.3. Воспроизводимость MCC = 0,199 и согласованность предсказаний при нескольких попытках обучения

Мы проверили, воспроизводит ли Conversational RuBERT результат MCC = 0,199 и пришли к выводу, что результат был получен случайно. Вторая модель, тренированная на FUN, в течение 12 эпох обучения на A_HDE, достигла максимального значения MCC = 0.147. Ответы моделей на 86,9% совпали (556 совпадений, 84 разных ответа).

Мы не нашли значимых различий в качестве распознавания текстов разных типов и текстов с коннектором в сет-апе или панчлайне.

Типы текстов	Red Light	Garden Path	Crossroad
Всего в тестовой части A_HDE	398	202	40

Совпадения предсказаний двух моделей Conv. RuBERT	345	180	32
Совпадения & правильные предсказания	188	123	19

Табл. 5. Количество текстов разных категорий. В данном случае неюмористическим текстам был присвоен тот же тип (RL, GP, C), который был у их юмористического варианта.

Позиция коннектора	Сет-ап	Панчлайн
Всего в тестовой части A_HDE	247	330
Совпадения предсказаний двух моделей Conv. RuBERT	450	640

Табл. 6. Количество текстов (шуток и не-шуток) с двумя возможными позициями коннектора.

В тестовом датасете оказался 91 инвариантов (пар 1–0 или троек 1–0–0), в которых хотя бы юмористический текст и хотя бы один неюмористический были угаданы верно, то есть они составляли $\frac{91}{315} \times 100\% = 28{,}9\%$ от всего количества инвариантов в тестовом датасете. Это число не свидетельствует, что некоторый признак в этих инвариантах сделал их более «распознаваемыми». Если считать, что вероятность правильно угадать любой из лейблов $p = 0{,}5$, то вероятность угадать правильно два лейбла в паре $p_2 = 0{,}5^2 = 0{,}25$, а в тройке $p_3 = \frac{C_3^2 + C_3^3}{0{,}5^{-3}} = \frac{3+1}{8} = 0{,}5$. Полученное число инвариантов находится в промежутке между ожидаемыми количествами угаданных инвариантов в паре и в тройке $0{,}25 < 0{,}289 < 0{,}5$.

По-видимому, единственное, с чем была корреляция в предсказаниях — среднее расстояние в токенах между юмористическим и неюмористическим текстами, противопоставленными в рамках инварианта (аналог расстояния Левенштейна, где вместо символов — токены, а вместо строк — тексты). Для не-шуток мы посчитали, сколько токенов было заменено, удалено или добавлено, чтобы получить данную не-шутку из шутки. Мы предположили, что у не-шуток, для которых модель правильно предсказала лейбл, среднее расстояние в токенах будет больше, чем у не-шуток с неугаданным лейблом. Эта гипотеза подтвердилась (однонаправленный t-теста Стьюдента, t-statistic = 2,42, p-value = 0,008, результат значим при p-value < 0,01).

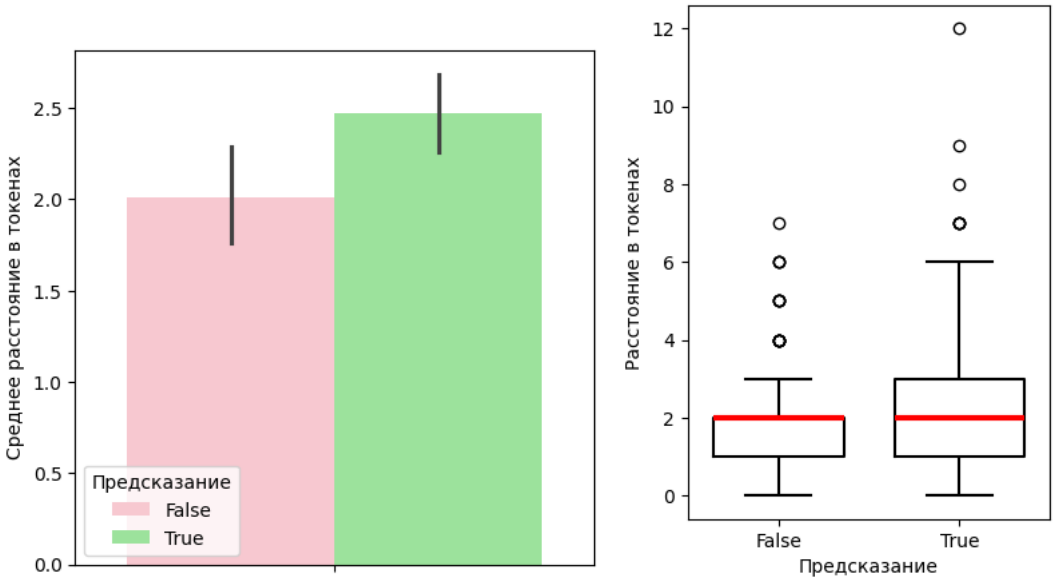

Рис. 7. Группы правильно и неправильно предсказанных неюмористических текстов.

4.4. Эксперименты с обучением на FUN (3 эпохи) и тестирование на A_HDE

Мы использовали полный датасет A_HDE для проверки тестирования моделей, обученных только на FUN. Conversational RuBERT показал MCC = 0.055, RuBERT показал MCC = 0.043. В ответах моделей не было каких-либо закономерностей в типе юмора, позиции коннектора и именованных сущностях в тексте.

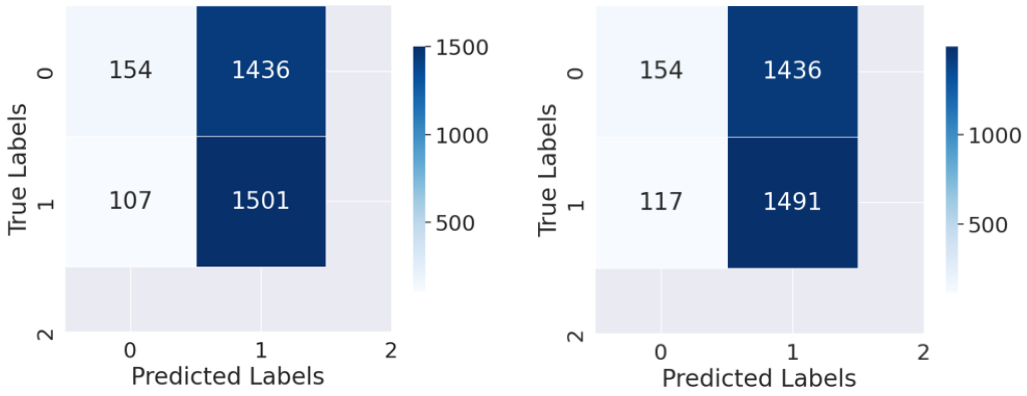

Рис. 8. (слева) Conversational RuBERT, confusion matrix.

Рис. 9. (справа) RuBERT, confusion matrix.

Также для сравнения приводим результаты RuBERT и Conversational RuBERT с рандомно инициализированными весами.

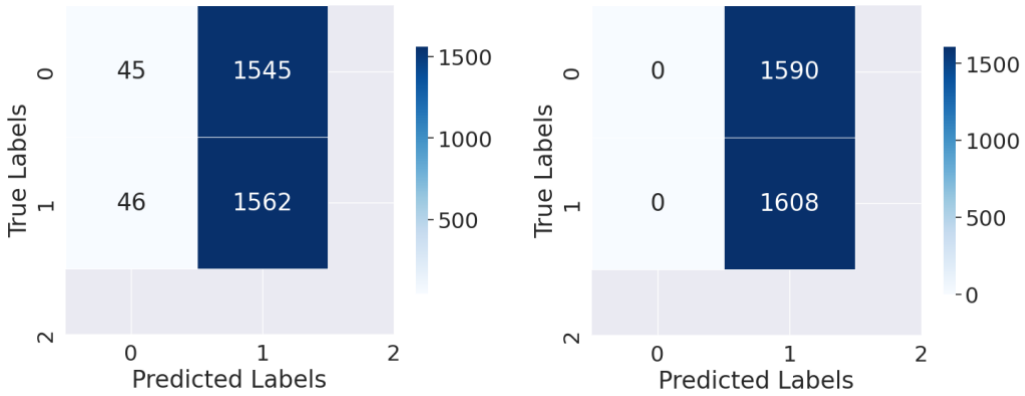

Рис. 10. (слева) Conversational RuBERT, confusion matrix. MCC = 0.

Рис. 11. (справа) RuBERT, confusion matrix. MCC = –0,001.

4.2.1. Предсказания модели Conversational BERT при лучшем MCC за все эксперименты

Мы подробнее рассмотрели предсказания модели Conversational BERT, обучившейся 3 эпохи на FUN и 7 эпох на A_HDE (на тестовой выборке MCC = 0,199). Нас интересовало, зависело ли качество предсказаний от типов шуток, позиции коннектора и побочных факторов.

Позиция коннектора и наличие именованных сущностей в тексте оказались незначимыми факторами. Мы обнаружили, что модель лучше распознавала шутки и не-шутки типа Garden Path, чем Red Light и Crossroads.

	True positive	False positive
GP	74	35
RL + C	150	117

Chi-square = 4,4, p-value = 0,036. Результат значим при p-value < 0,05.

	True negative	False negative
GP	65	28
RL + C	95	77

Chi-square = 5,42, p-value = 0,019. Результат значим при p-value < 0,05.

Наличие именованных сущностей не оказало статистически значимого влияния на предсказания модели.

4.4. Косинусная схожесть сет-апа и панчлайна.

В работе [Zakovorotnaia 2022] предлагается использовать косинусную близость между сет-апами и панчлайнами анекдотов, чтобы обнаружить переключение между двумя интерпретациями / фреймами / скриптами [Raskin 1984]). Евгения Заковоротная показывает, что косинусное расстояние эффективно, чтобы различать противопоставлений сет-ап vs. панчлайн в шутках и противопоставлений соседних предложений в художественных текстах.

Мы использовали предложенный метод, но вместо художественных текстов использовали неюмористические тексты из нашего датасета. Для создания эмбедингов и вычисления косинусной близости мы использовали токенизатор для Conversational RuBERT и модель Conversational RuBERT, обученную 3 эпохи на FUN. В такой вариации метод оказался неэффективен.

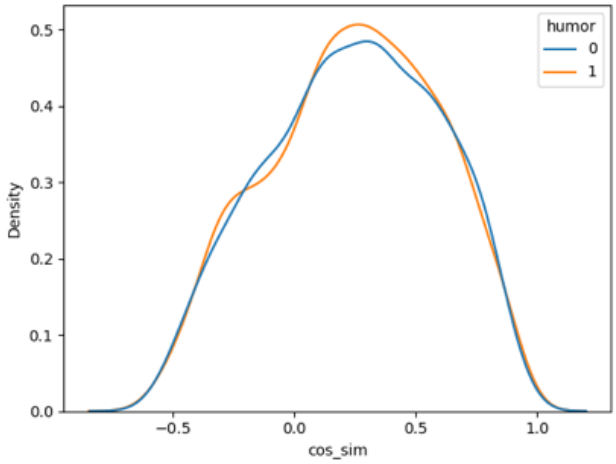

Рис. 12. Косинусное расстояние между сет-апом и панчлайном.

5. Выводы

Мы создали датасет A_HDE, состоящий из 1608 текстов юмористических текстов и 1590 неюмористических. В датасете присутствует разметка по нескольким параметрам, позволяющих формально описать юмор с точки зрения логики построения текста и лингвистических черт, а именно тип юмористического механизма garden path, red light, crossroad и позицию коннектора — выражения, позволяющего слушателю / читателю переключиться между двумя разными интерпретациями юмористического текста. Датасет может быть отфильтрован более строго, может быть использован в качестве источника для отбора шуток определённого типа. Также подготовлены рекомендации по пополнению датасета и превращению юмористического текста в неюмористический по методу минимальных инвариантов.

Эксперименты, в которых использовался датасет A_HDE, показали, что его объёма не хватает для полноценного обучения моделей BERT задаче humor detection и что модели, обучение на других датасетах не помогает научить модель распознавать юмористические механизмы. Обучение на датасетах, созданных с менее строгими правилами к неюмористическим примерам, может быть полезно для практических задач, но не позволяет сделать теоретических выводов о юмористических чертах и способностях моделей находить такие черты.

Мы считаем интересным результатом, что тип шуток garden path оказался более лёгким для распознавания в одном экспериментов. Причём, перфоманс Conversational RuBERT был лучше и на юмористических, и неюмористических примерах. Мы предполагаем, что этот тип юмора наиболее легко формализировать. Сет-ап всегда заставляет слушателя / читателя прийти к определённой интерпретации, которая обязательно оказывается неверной в панчлайне. Легкодоступность первой интерпретации может быть обусловлена за счёт частотности употребления определённых слов в контексте сет-апа, а контекст в панчлайне создаёт контраст с предыдущим контекстом. В шутках с механизмом garden path первая интерпретация может быть более и менее явной. Интересной задачей будет придумать метрику для определения её явности.

Об этике.

Юмористические тексты были собраны из открытых источников и могут содержать неэтичные высказывания.

Благодарности

Научному руководителю Дарье Александровне Рыжовой, доценту, преподавателю НИУ ВШЭ.

Альберту Корнилову. За предоставление видеокарты A100 для обучения модели.

Полине Офимкиной и Cristine Howes. За консультации.

Список литературы и источники

Аристотель 2000 — Риторика. Поэтика. М.: «Лабиринт», 2000.

Архипова 2003 — Архипова Александра Сергеевна. Анекдот и его прототип: генезис текста и формирование жанра. Автореферат диссертации на соискание ученой степени кандидата филологических наук. РГГУ, Москва 2003

Шмелева, Шмелев 2002 — Русский анекдот как текст и как речевой жанр. Е. Я. Шмелева · А. Д. Шмелев. Русский язык в научном освещении. 2002. № 2 (4), 194-210.

Рахилина, Плунгян 2009 — Рахилина Е. В., Плунгян В. А. Анекдот как конструкция // Известия РАН. Серия литературы и языка, 68, № 5, 2009.

Aarons Debra 2017 — Puns and Tacit Linguistic Knowledge
ByDebra Aarons. Attardo, S. (Ed.). (2017). The Routledge Handbook of Language and Humor (1st ed.). Routledge. 80 - 95

Aljared,A. (2017).The isotopy-disjunction model. In S. Attardo (Ed.), The Routledge handbook of language and humor (pp. 64–79). New York: Routledge.

Attardo 1994 — Salvatore Attardo. Linguistic theories of humor. Berlin, Germany: Mouton de Gruyter.

Blinov et al: Large Dataset and Language Model Fun-tuning for Humor Recognition // ACL, (2019).

Chiruzzo 2021 — Chiruzzo, L., Castro, S., Góngora, S., Rosa, A., Meaney, J. Mihalcea, R. (2021). Overview of HAHA at IberLEF 2021: Detecting, Rating and Analyzing Humor in Spanish. Procesamiento de Lenguaje Natural. 67. 257-268. 10.26342/2021-67-22.

Devlin et al. 2018 — J. Devlin, M.-W. Chang, K. Lee, K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding // Proceedings of The 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics, 2018.

Dynel, M. 2012 — “Garden-paths, red lights and crossroads: On finding our way to understanding the cognitive mechanisms underlying jokes.” Israeli Journal of Humor Research: An International Journal 1: 6-28.

Forabosco 2008 — "Cognitive aspects of the humor process: the concept of incongruity" HUMOR, vol. 5, no. 1-2, 1992, pp. 45-68. https://doi.org/10.1515/humr.1992.5.1-2.45

Grice 1975 — "Logic and conversation". In Cole, P.; Morgan, J. (eds.). Syntax and semantics. Vol. 3: Speech acts. New York: Academic Press. pp. 41–58.

Hockett, C.F. (1973). Studies in linguistics in Honor of George L. Trager. Mouton: The Hague.

Inácio, M., Wick-pedro, G., Oliveira, H. G. (2023). What do Humor Classifiers Learn? An Attempt to Explain Humor Recognition Models. In Proceedings of the 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, pages 88–98, Dubrovnik, Croatia. Association for Computational Linguistics.

Kao et al. 2015 — J. T. Kao., R. Levy, N. D. Goodman. A Computational Model of Linguistic Humor in Puns. Cognitive Science, No. 1-16, 2016. 1270–1285

Krikmann 2006 — Eesti Kirjandusmuuseum. Folklore: Electronic Journal of Folklore. 2006. 33. 27-58

Kuratov, Arkhipov 2019 — Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language.

Larkin-Galiñanes 2017 — Cristina Larkin-Galiñanes. An Overview of Humor Theory. The Routledge Handbook of Language and Humor. Salvatore Attardo (ed.) New York: Routledge. P. 49–63.

McGhee 1972 — In The Psychology of Humor: Theoretical Perspectives and Empirical Issues, edited by Jeffrey H. Goldstein and Paul E. McGhee, 81-100. New York: Academic Press, 1972.

Peyrard et al. 2021 — Peyrard, M., Borges, B., Gligorić, K., West, R. (2021). Laughing Heads: Can Transformers Detect What Makes a Sentence Funny? Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence. Montreal, 19-27 August 2021. International Joint Conferences on Artificial Intelligence.

Rada Mihalcea and Carlo Strapparava. 2005. Making Computers Laugh: Investigations in Automatic Humor Recognition. In Proceedings of Human Language Technology Conference and

Conference on Empirical Methods in Natural Language Processing, pages 531–538, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Raskin, V. (1985). Semantic Mechanisms of Humor. Dordrecht, Boston, MA, Lancaster: D. Reidel.

Ren, C., Guo, Z., Zhang, P., and Gao, Y. (2024). Humor detection using deep learning in 10 years: A survey. Revista Internacional de Métodos Numéricos para Cálculo y Diseño en Ingeniería. Vol. 40, (1), 4. URL https://www.scipedia.com/public/Ren_et_al_2023b

Suls, Jerry M. “A two-stage model for the appreciation of jokes and cartoons.” In The Psychology of Humor: Theoretical Perspectives and Empirical Issues, edited by Jeffrey H. Goldstein and Paul E. McGhee, 81-100. New York: Academic Press, 1972.

Weller, O. and Seppi, K. (2019). Humor detection: A transformer gets the last laugh. 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), pages 3621–3625. Hong Kong, China. Association for Computational Linguistics.

Westbury et al. 2016 — Chris Westbury, Cyrus Shaoul, Gail Moroschan, Michael Ramscar. Telling the world’s least funny jokes: On the quantification of humor as entropy, Journal of Memory and Language. Volume 86, 2016. P. 141–156.

Winters, T., & Delobelle, P. (2020). Dutch Humor Detection by Generating Negative Examples. ArXiv, abs/2010.13652.

Ziv 1984 — Ziv, Avner. 1984. Personality and Sense of Humor. New York: Springer.