ИЗДАТЕЛЬСТВО ГРАМОТА

Филологические науки. Вопросы теории и практики
Philology. Theory & Practice

ISSN 2782-4543 (online)
ISSN 1997-2911 (print)

2024. Том 17. Выпуск 4 | 2024. Volume 17. Issue 4
Материалы журнала доступны на сайте (articles and issues available at): philology-journal.ru

RU

Использование машинного обучения для тематической разметки текстовых материалов корпуса устной речи

Погодаева Е. Н.

Аннотация. Цель исследования состоит в выявлении эффективности тезаурусного метода для формирования списка тематических классов при использовании машинного обучения для тематической классификации текстовых материалов социолингвистических интервью. В статье рассматриваются возможности применения машинного обучения в тематической разметке материалов лингвистического корпуса. Политематичность анализируемого материала обусловлена его жанровой принадлежностью к диалогической речи. Иерархическая структура тем, выявленная в результате предварительного интроспективного анализа текстов, может быть описана с помощью тезауруса. Обсуждаются результаты применения метода машинного обучения без учителя с использованием двух наборов названий тематических классов: списка тем, задействованного при ручной разметке текстов, и расширенного списка микротем, названия которых были отобраны из тезауруса русского языка. Научная новизна работы состоит в том, что впервые предложен тезаурусный метод подбора тематических лейблов для zero-shot классификации слабоструктурированных текстов на русском языке. Полученные результаты показали, что использование более подробного лексического описания для тематических классов даёт улучшение результата классификации.

EN

Using machine learning for the topic annotation of oral speech corpus texts

E. N. Pogodaeva

Abstract. The research aims to determine the effectiveness of the thesaurus method for forming a list of topic classes when using machine learning for the topic classification of text materials of sociolinguistic interviews. The paper considers the potential of using machine learning in the topic annotation of linguistic corpus materials. The polytopical nature of the analyzed material is due to its genre belonging to dialogical speech. The hierarchical structure of the topics, identified as a result of a preliminary introspective analysis of the texts, can be described using a thesaurus. The results of using the unsupervised machine learning method are discussed involving two sets of topic class names: a list of topics used in manual text annotation and an extended list of micro-topics whose names were selected from a Russian language thesaurus. The paper is novel in that it is the first to propose the thesaurus method for selecting topic labels for the zero-shot classification of weakly structured Russian texts. The research findings show that using a more detailed lexical description for topic classes improves the classification result.

Введение

Процесс сбора и разметки данных лингвистического корпуса предполагает использование информационных технологий на всех этапах работы. Отдельные виды корпусной разметки (например, морфологическая, синтаксическая, словообразовательная) в настоящий момент полностью автоматизированы, однако в случаях с другими уровнями языка, для которых производится аннотирование, при использовании средств автоматической обработки текстов можно столкнуться с рядом проблем. К таким случаям относится задача тематической разметки материалов неподготовленной диалогической разговорной речи, так как ввиду политематичности как одного из основных свойств разговорной диалогической речи один текст может содержать несколько тем, а смена темы в границах одного диалогического единства может происходить несколько раз.

Актуальность темы исследования обусловлена потребностью адаптации существующих средств автоматической обработки текстов для разметки корпусов текстов русского языка. Применение машинного обучения

Научная статья (original research article) | https://doi.org/10.30853/phil20240186

с использованием предобученной языковой модели позволит размечать большие объёмы текстовых данных без привлечения дополнительных ресурсов для разметки достаточной по объёму обучающей выборки, требуемой для применения классических методов машинного обучения с учителем, и значительных временных затрат, необходимых для ручной разметки.

Для достижения поставленной цели исследования в статье решены следующие задачи: 1) проанализировать тематическую структуру социолингвистических интервью; 2) составить список тематических классов для алгоритма тематической классификации; 3) применить zero-shot классификатор с двумя наборами тематических классов.

Выбор методов исследования обусловлен целью исследования и совокупностью поставленных задач. Для предварительного выделения тем текстов на этапе ручной разметки был использован метод лингвистической интроспекции. Названия тематических классов были подобраны на основе тезаурусного метода. Эксперименты по автоматической тематической классификации были произведены с помощью машинного обучения с использованием предобученной языковой модели.

Материалом исследования являются 25 социолингвистических интервью общим объёмом 214139 словоупотреблений с предварительно выполненной ручной тематической разметкой. Интервью были взяты из находящегося в разработке корпуса русской устной речи тюркско-русских билингвов RuTuBiC. Отбор текстов производился методом случайной выборки.

Теоретической базой исследования послужили работы, посвящённые методам и инструментам автоматической разметки лингвистических корпусов (Захаров, Богданова, 2020; Ляшевская, 2016). Вопросы построения и применения тезаурусов рассматриваются в работе Н. В. Лукашевич (2010), в которой представлены особенности структуры тезауруса русского языка РуТез и типовые примеры описания отношений в тезаурусе. Также данное исследование опирается на опыт применения тезауруса в процессе корпусного моделирования (Баранов, Добровольский, 2021), в частности – на опыт подготовки словника тезауруса. Авторы публикаций, посвящённых применению предобученных языковых моделей для классификации текстов на высокоресурсных языках (Rothman, 2024; Singh, 2023; Wang, Pang, Lin, 2023; Plaza-del-Arco, Nozza, Hovy, 2023), отмечают, что выбор zero-shot метода машинного обучения позволяет использовать предобученную языковую модель для предсказания классов, которые не были задействованы при обучении модели. Рассматривая методы решения задачи классификации в условиях ограниченных временных и вычислительных ресурсов, авторы исследований в области многоклассовой тематической классификации (Bhambhoria, Chen, Zhu, 2023; Zhang, Yang, Xu et al., 2024) отмечают положительное влияние применения расширенной иерархии тематических рубрик на результаты классификации с использованием предобученных больших языковых моделей.

Практическая значимость исследования состоит в возможности применения его материалов и результатов при разметке лингвистических корпусов. Полученные результаты также могут быть использованы при разработке учебных курсов по корпусной лингвистике и автоматической обработке естественного языка.

Обсуждение и результаты

Тексты корпуса размечены по шести параметрам: морфология, отклонения от речевого стандарта, коррекция отклонений от речевого стандарта, дискурс, тип речи, жанр и тема. В настоящий момент средства автоматической обработки текстов применяются только для морфологической разметки. Морфологическая разметка производится с помощью морфологического анализатора русского языка Mystem. Разметка по остальным параметрам осуществляется вручную. Подробное описание типологических параметров корпуса и этапов работы по его созданию представлено в серии публикаций коллектива создателей корпуса (Резанова, 2017, с. 105-118; 2019, с. 127-140).

Текстовые материалы корпуса хранятся в закрытой базе данных (Артёменко Е. Д., Буб А. С., Васильева А. В., Душейко А. С., Машанло Т. Е., Нагель О. В., Резанова З. И., Сафиуллина Е. Ш., Степаненко А. А., Темникова И. Г. Свидетельство о государственной регистрации базы данных № 2019620803 Российская Федерация. Бимодальный корпус устной речи жителей Южно-Сибирского региона: № 2019620689: заявл. 07.05.2019: опубл. 22.05.2019; заявитель Федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский Томский государственный университет»). Каждое интервью маркировано шифром, включающим в себя обозначение материнского языка билингва (Sh – для шорского языка, T – для татарского языка, H – для хакасского языка), порядковый номер респондента и порядковый номер интервью с респондентом.

Несмотря на то, что содержание диалогов в целом соответствует структуре социолингвистической анкеты (Казакевич, 2015) и модифицированной участниками проекта анкеты языкового опыта билингва (Marian, Blumenfeld, Kaushanskaya, 2007, p. 940-967), порядок следования вопросов и объём их обсуждения разнится от респондента к респонденту. В зависимости от возраста, пола и социального опыта респондента освещение некоторых тем может варьироваться в объёме. Для воспроизведения ситуации непринуждённого общения интервьюеры могли задавать уточняющие вопросы или вставлять собственные короткие реплики. Также имел значение факт знакомства участников коммуникации до записи диалога. В отдельных случаях коммуниканты могли состоять в родственных или дружественных отношениях, поэтому такие интервью имеют свободный характер.

Тематическая разметка корпуса RuTuBiC имеет иерархическую структуру. После предварительного анализа текстов было выделено две больших группы тем: «мир» и «мнение о мире». Темы группы «мир» делятся на три подгруппы: «человек», «среда», «событие». Темы группы «мнение о мире» делятся на две подгруппы в зависимости от количества участников диалога, выражающих мнение: «мнение» (один участник, тег TOpinion), «обмен мнениями» (два и более участников, тег TOpinion exchange). Тематическая подгруппа «человек» включает в себя три темы: «рассказ о себе» (тег TSelf-characterization), «характеристика человека» (тег THuman characteristics), «семья» (тег TFamily). Тематическая подгруппа «среда» представлена темами «образ жизни» (тег TMode of life), «окружающая среда» (тег TNatural environment), «социальное окружение» (тег TSocial environment), «этническая культура» (тег TEthnic culture). Подгруппа тем «событие» включает в себя темы «событие в жизни» (тег TEvent in life), «период жизни» (тег TPeriod of life), «факт о событии» (тег TFact). При разметке текстов маркируются только отдельные темы без указания тематических групп. Ввиду политематичности как одного из основных свойств диалогической речи во избежание слишком дробного маркирования было принято решение использовать более одного тематического тега для фрагмента текста. То же относится к тематической группе «мнение», теги «мнение» и «обмен мнениями» было рекомендовано использовать в паре с одним из тегов тематической группы «мир». Также отдельно выделяются атематические фрагменты (тег TCurrent situation), содержание которых относится не к содержанию диалога, а к обстоятельствам его протекания.

Так как объём собранного языкового материала в корпусе RuTuBiC не достаточно большой для применения методов машинного обучения с учителем, а уже выполненная ручная тематическая разметка имеет высокую степень субъективности, для тематической классификации была выбрана zero-shot классификация (Song, Upadhyay, Peng et al., 2019, p. 133-150) – метод машинного обучения, не требующий обучающей выборки, при котором предобученная языковая модель может определить принадлежность объектов к классам, которых не было в процессе обучения модели. Такой подход позволяет модели обобщать свои знания и классифицировать объекты на основе общего понимания классов даже в случае отсутствия обучения на этих классах.

Задача сегментации текстов выходит за рамки тематической классификации, поэтому для проведения двух экспериментов с разными наборами тематических классов было принято решение разделить диалоги на сегменты в соответствии с ручной разметкой, приняв во внимание не принадлежность к конкретной теме, а факт её смены.

Эксперименты с двумя разными наборами тематических классов были проведены на двадцати пяти текстах, которые были уже размечены вручную, чтобы сравнить результаты ручной и автоматической тематической разметки. Для проведения обоих экспериментов была произведена предварительная обработка текстов. Из текстов были удалены все небуквенные символы, затем тексты были лемматизированы и очищены от стоп-слов.

Классификация осуществлялась с помощью transformers, библиотеки языка программирования Python. Для получения предсказаний была использована предобученная языковая модель с Hugging Face Hub joeddav/xlm-roberta-large-xnli. Так как, согласно изначальной идее создателей корпуса RuTuBiC, каждый сегмент текста может быть отнесён к более чем одной теме, была использована опция multi_label для многоклассовой классификации.

В первом эксперименте модели был передан список названий тем, применяемых авторами корпуса для ручной разметки. Темы «рассказ о себе» и «характеристика человека» были объединены в одну тему «характеристика человека», а темы «событие» и «факт о событии» были объединены в одну тему «событие». Атематические фрагменты в данном эксперименте не были выделены в отдельную категорию. Таким образом, были взяты следующие названия тематических классов: «характеристика человека», «семья», «образ жизни», «окружающая среда», «социальное окружение», «этническая культура», «событие», «период жизни».

Интерпретация результатов классификации производилась автором статьи в связи с высокой субъективностью исходной ручной разметки. Приведём пример присвоения тематических классов в случайно выбранном диалоге (см. Таблицу 1).

Таблица 1. *Сравнение тем, присвоенных вручную и автоматически*

Тема, присвоенная вручную	**Темы, присвоенные автоматически в порядке убывания вероятности**		
TCurrent situation			
TSelf-characterization	характеристика человека	период жизни	событие
TSelf-characterization: TSocial environment	период жизни	характеристика человека	социальное окружение
TSelf-characterization	период жизни	событие	образ жизни
TSocial environment	социальное окружение	период жизни	событие
TSocial environment: TEthnic culture	характеристика человека	социальное окружение	этническая культура
TSocial environment	период жизни	характеристика человека	событие
TEthnic culture	период жизни	этническая культура	характеристика человека
TFamily:TEthnic culture	семья	период жизни	этническая культура
TSelf-characterization	событие	характеристика человека	период жизни
TMode of life	характеристика человека	окружающая среда	период жизни
TEthnic culture: TSocial environment	характеристика человека	период жизни	социальное окружение
TEthnic culture	событие	период жизни	социальное окружение

Для интерпретации результатов были взяты три самых вероятных тематических класса. Для четырёх фрагментов текста тема, определённая классификатором, не совпала с темой, определённой человеком. Автоматически определённый тематический класс совпал с ручной разметкой в трёх случаях, для ещё пяти фрагментов тема, полученная в результате классификации, совпала с исходной темой только на втором или третьем по вероятности тематическом классе. Атематический фрагмент в рассматриваемом диалоге состоял из одного предложения, в котором не осталось слов после этапа предобработки, поэтому для него отсутствуют предсказанные тематические лейблы в обоих экспериментах.

Второй эксперимент был произведён с названиями тематических классов, отобранных из тезауруса русского языка RuWordNet. Использование тезауруса обусловлено тем, что данный источник лексического описания может быть применён для тематической разметки разной степени детализации. Каждая тема из первого эксперимента была расширена списком микротем или заменена на синонимичное описание, как в случае с темами «образ жизни» и «период жизни». В Таблице 2 сопоставлены два перечня названий тематических классов для первого и второго экспериментов соответственно.

Таблица 2. *Сравнение тем, присвоенных вручную и автоматически*

Названия тематических классов для первого эксперимента	**Названия тематических классов для второго эксперимента**
характеристика человека	человеческая личность, линия поведения
семья	семейная среда, наличие детей, родственник
образ жизни	жизненный уклад
окружающая среда	среда жизнедеятельности, антропогенная деятельность, защита окружающей среды, климатическо-природные условия, окружающая природная среда, дикая природа, животный и растительный мир
социальное окружение	круг общения, социальная среда, социокультурная среда
этническая культура	этническая культура, этническая принадлежность, языковая среда, родной язык, тюркская группа языков, хакасский язык, татарский язык, шорский язык, фольклор, музыкальный фольклор, танцевальный фольклор, народное предание
событие	происшествие, случай, описание фактов, ряд событий, мероприятие
период жизни	временной период

В Таблице 3 представлены результаты классификации с названиями тем из тезауруса для того же текста.

Таблица 3. *Сравнение тем, присвоенных вручную и автоматически*

Тема, присвоенная вручную	**Темы, присвоенные автоматически в порядке убывания вероятности**		
TCurrent situation			
TSelf-characterization	человеческая личность	родственник	среда жизнедеятельности
TSelf-characterization: TSocial environment	языковая среда	социальная среда	среда жизнедеятельности
TSelf-characterization	языковая среда	тюркская группа языков	временной период
TSocial environment	социальная среда	социокультурная среда	временной период
TSocial environment: TEthnic culture	тюркская группа языков	человеческая личность	языковая среда
TSocial environment	хакасский язык	языковая среда	тюркская группа языков
TEthnic culture	языковая среда	этническая принадлежность	родной язык
TFamily:TEthnic culture	родственник	шорский язык	народное предание
TSelf-characterization	временной период	среда жизнедеятельности	линия поведения
TMode of life	животный и растительный мир	человеческая личность	родственник
TEthnic culture: TSocial environment	шорский язык	этническая принадлежность	происшествие
TEthnic culture	хакасский язык	шорский язык	языковая среда

По результатам второго эксперимента тема не была определена верно в трёх случаях, однако самый вероятный тематический класс совпадал с ручной разметкой в семи случаях, и ещё в двух случаях тема, совпадающая с исходной темой, была второй или третьей по вероятности.

Несмотря на то, что количество неверно интерпретированных названий тематических классов в двух экспериментах отличается незначительно, во втором случае больше верных ответов за счёт выделенных микротем, использованных в качестве списка тематических классов. Сравним ручную разметку и результаты классификации в двух проведённых экспериментах для отдельно взятого текстового фрагмента (см. Таблицу 4).

И вот вы на шорском языке сейчас немножко говорите, да? Да, немножко. Понимаете, что говорят? Понимаю не всегда. Не всегда? Не всегда понимаю, ну вот… А вам что проще? Слушать, понимать или разговаривать на шорском. Мне, наверно, больше слушать. Сложнее, да? Проще, да? Проще, да, да, труднее разговаривать. Потому что… ну вот так. Угу. А читать на шорском языке вы можете? На шорском могу читать языке, потому что у меня, говорю, казахский изучал, грамматика одна, там какие-то есть маленько разницы в произношении,

там вообще незначительные, а… А писать на шорском языке можете? Ну, тоже, если что-то такое, простое… А бывает, что иногда вот какая-то… (Артёменко, Буб, Васильева и др., 2019, Sh0024_001).

Таблица 4. *Сравнение ручной разметки и результатов классификации в двух проведённых экспериментах*

Тема, присвоенная вручную	Эксперимент 1	Эксперимент 2
TEthnic culture	событие, период жизни, социальное окружение	хакасский язык, шорский язык, языковая среда

Приведённый фрагмент посвящён обсуждению разных аспектов использования шорского языка респондентом и не включает упоминаний ни одной из тем, полученных в результате первого эксперимента. Тема «этническая культура» была определена с вероятностью 0.52 (не включена в перечень в Таблице 1). Такой результат можно объяснить тем, что в приведённом фрагменте текста в фокусе обсуждения находится именно язык, а не культура. Во втором эксперименте модель определила, что речь идёт именно о языке, хотя и в этом, и в других текстах включение в список тематических лейблов названий конкретных языков не повлияло на точность определения темы. В обоих экспериментах тема «образ жизни» не была определена верно, однако в обоих экспериментах использовался только один вариант названия темы.

Да. Ну а отец, рыбалка, охота, вот это постоянно. Ну также собирательство, как можно так назвать. Это колба, черемша, наверно, знаете? Папоротники, грибы, ягоды, вот это вот, орех кедровый, это вот всё. Пушнину тоже добывает, но это тоже там… соболь… соболей. Вот. Такие вот дела (Артёменко, Буб, Васильева и др., 2019, Sh0024_001).

В тексте есть упоминание растительного и животного мира (окружающая среда), а также упоминание родственника (характеристика человека), однако полученные тематические классы не описывают типовые образцы протекания жизни.

Таким образом, анализ приведённых примеров подтверждает эффективность тезаурусного метода для формирования списка тематических классов при использовании предобученной языковой модели для тематической классификации текстов.

Заключение

Выполненное экспериментальное исследование позволяет сделать следующие выводы:

1. Тематическое членение социолингвистического интервью ввиду своей жанровой принадлежности к диалогической речи имеет иерархическую структуру.

2. Иерархия тематических классов в зависимости от цели и выбранного метода разметки может иметь разную глубину, оставаясь при этом неизменной на макроуровне.

3. Тезаурусное описание тематических классов положительно влияет на результат zero-shot классификации при отсутствии возможности применения классических методов машинного обучения с учителем.

Перспективы дальнейшего исследования состоят в использовании представленных результатов в разметке корпуса русской устной речи тюркско-русских билингвов RuTuBiC. Для получения более точных результатов планируется провести аналогичный эксперимент со сбалансированным количеством микротем для каждого тематического класса. Также выбранный метод машинного обучения планируется использовать для тем группы «мнение о мире».

Источники | References

1. Баранов А. Н., Добровольский Д. О. Корпусная модель идиостиля Достоевского. М.: ЛЕКСРУС, 2021.
2. Захаров В. П., Богданова С. Ю. Корпусная лингвистика. СПб.: Изд-во С.-Петерб. ун-та, 2020.
3. Казакевич О. А. О принципах построения функциональной типологии малых языков (на материале малых автохтонных языков Сибири и Дальнего Востока) // Функциональное развитие языков в полиэтнических странах мира (Россия – Вьетнам): материалы международного круглого стола. М.: Азбуковник, 2015.
4. Лукашевич Н. В. Тезаурусы в задачах информационного поиска. М., 2010.
5. Ляшевская О. Н. Корпусные инструменты в грамматических исследованиях русского языка. М.: Издательский дом ЯСК; Рукописные памятники Древней Руси, 2016.
6. Резанова З. И. Корпус устной речи русско-тюркских билингвов Южной Сибири: разметка отклонений от речевого стандарта // Вопросы лексикографии. 2019. № 15.
7. Резанова З. И. Подкорпус устной речи русско-тюркских билингвов Южной Сибири: типологически релевантные признаки // Вопросы лексикографии. 2017. № 11.
8. Bhambhoria R., Chen L., Zhu X. A Simple and Effective Framework for Strict Zero-Shot Hierarchical Classification // arXiv. 2023. Art. 2305.15282. https://doi.org/10.48550/arXiv.2305.15282
9. Marian V., Blumenfeld H. K., Kaushanskaya M. The Language Experience and Proficiency Questionnaire (LEAP-Q): Assessing Language Profiles in Bilinguals and Multilinguals // Journal of Speech, Language, and Hearing Research. 2007. Vol. 50 (4).

10. Plaza-del-Arco F., Nozza D., Hovy D. Wisdom of Instruction-Tuned Language Model Crowds. Exploring Model Label Variation // arXiv. 2023. Art. 2307.12973. https://doi.org/10.48550/arXiv.2307.12973.
11. Rothman D. Transformers for Natural Language Processing and Computer Vision. Birmingham: Packt Publishing, 2024.
12. Singh J. Natural Language Processing in the Real World: Text Processing, Analytics, and Classification. 1st ed. N. Y.: Chapman and Hall, 2023.
13. Song Y., Upadhyay S., Peng H., Mayhew S., Roth D. Toward Any-Language Zero-Shot Topic Classification of Textual Documents // Artificial Intelligence. 2019. Vol. 274.
14. Wang Z., Pang Y., Lin Y. Large Language Models Are Zero-Shot Text Classifiers // arXiv. 2023. Art. 2312.01044. https://doi.org/10.48550/arXiv.2312.01044
15. Zhang Y., Yang R., Xu X., Xiao J., Shen J., Han J. TELEClass: Taxonomy Enrichment and LLM-Enhanced Hierarchical Text Classification with Minimal Supervision // arXiv. 2024. Art. 2403.00165. https://doi.org/10.48550/arXiv.2403.00165

Информация об авторах | Author information

RU

Погодаева Елена Николаевна[1]
[1] Томский государственный университет

EN

Elena Nikolaevna Pogodaeva[1]
[1] Tomsk State University

[1] *elenanikolaevnapogodaeva@gmail.com*

Информация о статье | About this article

Дата поступления рукописи (received): 20.02.2024; опубликовано online (published online): 25.04.2024.

Ключевые слова (keywords): лингвистический корпус; машинное обучение; тематическая классификация; разметка данных; диалогическая речь; linguistic corpus; machine learning; topic classification; data annotation; dialogical speech.