НЕЧЕТКАЯ ЛОГИКА НА БАЗЕ ЯЗЫКОВЫХ МОДЕЛЕЙ: ПОСТРОЕНИЕ ГИБРИДНОЙ СИСТЕМЫ КЛАССИФИКАЦИИ ДИАЛЕКТОВ И ОПРЕДЕЛЕНИЯ ДИАЛЕКТИЗМОВ

Миндубаев Артур
Аспирант Академии Наук Республики Татарстан,
Институт прикладной семиотики.
E-mail: a.mindubaev@bk.ru

Аннотация: В данной работе поставлена цель создание информационной системы классификации диалектов и определения диалектизмов с использованием нейросетевой модели и механизма логического вывода на базе правил нечёткой логики. Данная система не имеет прямых аналогов для русского языка, что подчёркивает актуальность данной работы.

Для решения поставленной задачи использовался язык Python с библиотеками для обработки естественного языка, обучения нейронной сети, работы с трансферной моделью Google Bert и создания базы правил нечёткой логики для модели Мамдани.

В научно-исследовательской работе был выполнен анализ диалектологической предметной области и поиск методов решения задач диалектологии с помощью методов обработки естественного языка и машинного обучения. Автором была предложена архитектура нейросетевой модели для классификации диалектов и сформированы основные концепции алгоритма определения диалектизмов. Следующим шагом было произведено обучение нейросетевой модели на основе Bert и реализован экспериментальный алгоритм для поиска диалектизмов.

Ключевые слова: классификация диалектов, нейронные сети, LM, Bert, нечёткая логика, модель Мамдани.

Abstract: This work aims to create an information system for classifying dialects and identifying dialectics using a neural network model and a logical inference mechanism based on the rules of fuzzy logic. This system has no explicit analogues for the Russian language, which emphasizes the relevance of this work.

To solve the problem, the Python programming language was used with libraries for natural language processing, neural network training, Google Bert model usage and base of fuzzy logic rules creation for the Mamdani model.

The research work involved an analysis of dialectology field and a search for methods using natural language processing and machine learning methods for dialect tasks. The author offered the architecture of the neural network model for the dialect classification and formed the basic concepts of the dialethism's determination algorithm. The next step was a neural network model training on Bert outputs and dialethism's determination experimental algorithm implementation.

Keywords: dialect classification, neural network, LM, Bert, fuzzy logic, Mamdani model.

Вступление

Современная форма распространённых естественных языков зачастую имеет установленный набор норм и правил. Такая форма языка называется литературной или

стандартной. Однако стандартизация естественных языков – это достижение нескольких последних столетий. В то же время территориальная, экономическая или социальная обособленность групп людей, говорящих на одном и том же языке, неизбежно приводила к характерным отклонениям и порождала диалекты. Степень отличия диалектов сильно варьируется в зависимости от условий, из-за чего носители диалектов могут, как испытывать незначительные трудности при общении, так и полностью потерять возможность понимать друг друга. Поэтому диалектическая экспертиза текстов может быть крайне сложной и кропотливой работой.

Несмотря на это, в результате анализа предметной сферы и поиска готовых коммерческих решений сложилось мнение, что набор инструментов для экспертов в диалектологии скуден, и всё становится ещё хуже, если речь идёт об экспертных системах для первичной экспертизы текста. И это не удивительно, так как для построения эффективной экспертной системы придётся использовать самые современные методы обработки естественных языков (NLP), связанные, в первую очередь, с машинным обучением, и, в частности, с нейронными сетями и трансферным обучением. Данная же работа посвящена созданию нейросетевой модели для русского языка.

Анализ существующих методов и подходов

Задачу определения диалекта и поиска диалектизмов можно отнести к области понимания естественного языка. Так, для задачи определения диалекта системе потребуется определить признаки текста, присущие определённому диалекту, то есть классифицировать текст.

Аналогичным образом можно поступить с задачей поиска диалектизмов. Для её решения потребуется выделить из текста всех потенциальных кандидатов и произвести их классификацию.

Существующие методы определения диалектов на основе языковых моделей

Большинство систем идентификации языка работают с речью [A. Etman, A. A. L. Beex, 2015; 220]. Так, при фонотактическом моделировании используют два популярных подхода распознавание звуков на основе языковой модели (PRLM) и параллельное распознавание звуков на основе языковой модели (PPRLM). В обоих подходах языковые модели оценивают речевую транскрипцию на принадлежность к определённому диалекту, то есть производят классификацию.

Задачу определения диалекта в тексте зачастую сводят к задаче определения языка. В работе [Jauhiainen T, Lindén K, Jauhiainen H., 2019; 561] рассматривается метод HeLI и HeLI 2.0. HeLI – это метод обучения с учителем, использующий языковые модели и схожий с методом наивного Байеса. В статье предлагался новый способ улучшения метода HeLI 2.0 с помощью адаптации языковой модели. Адаптация выполняется итеративно, путём пополнения словаря модели с наивысшей оценкой уверенности n-граммами из тестовых данных.

Языковые модели для задачи классификации

Самым быстроразвивающимся и многообещающим подходом обработки языков являются языковые модели, обучаемые на больших корпусах текстов [Mitchell, Melanie, David C. Krakauer, 2024]. Одной из таких моделей в открытом доступе является Google Bert. Данный трансформер по оценкам GLUE достигает, а местами и превосходит человеческие показатели. Bert способен решать разные задачи классификации текста [Sun, C., Qiu, X., Xu,

Y., Huang, X., 2019]: определение грамматической корректности (CoLa), определение истинности логического вывода (MNLI), определение парафразов (MRPC), определение ответа на вопрос (QNLI) и определение тождественных вопросов (QQP). Для адаптации модели под конкретную задачу классификации обычно используют дополнительный слой нейронной сети прямого распространения.

Модели нечёткой логики

Нечёткая логика является мощным инструментом реализации систем принятия решения на базе правил с 1970-х годов [Zimmermann, H-J., 2010; 317]. Заде в своей работе [Zadeh, 1973; 28] впервые предложил концепцию контроллера на базе логических правил.

Фактически финальный этап дефазификации можно рассматривать как классификацию предлагаемого решения на основе входных значений переменных и системы логических правил. Примером системы классификации на основе системы правил нечёткой логики является [McBratney, Alex B., Adrian W. Moore, 1985; 165]. В данной работе авторы пытались реализовать климатическую классификацию.

В основе контроллера нечёткой логики лежит механизм логического вывода. Самым популярным механизмом логического вывода является модель Мамдани [Iancu, Ion, 2012; 325]. Модель Мамдани проста для вычисления, так как использует функции минимума и максимума для агрегации правил. Всё это позволяет создавать большие системы правил для принятия решений.

Классификация диалектов

Основой нейросетевой модели системы классификации диалектов будет модель обработки естественных языков Google Bert модификации Large, обученной компанией SberDevices для своей сисемы «Салют» преимущественно на русском корпусе текстов [Д. Антюхов, 2020]. Таким образом, словарь модели составляет 120138 уникальных токенов.

Для задачи классификации понадобится вектор скрытого состояния токена [CLS] размерности 1024. Он пропускается через линейный слой нейронной сети (pooler) с tanh функцией активации. Данный слой использовался для задачи предсказания следующего предложения во время первичного обучения модели. Именно выход pooler-слоя размерностью 1024 и станет входом для последнего слоя нейронной сети (head) для тонкой настройки под задачу классификации диалектов. Данный слой тонкой настройки был обучен на подготовленных данных подробнее о которых будет рассказано в разделе 5

В качестве финального слоя классификатора нейронной сети будет однослойная нейронная сеть прямого распространения размерностью (1024,3) с функцией активации Relu. Так как требуется провести классификацию по трём диалектам (северный, среднерусский и южный) будет использовала функция потерь кросс-энтропии. В качестве метода оптимизации был выбран Adam (Adaptive moment estimation).

Однако учитывая небольшую выборку обучающих примеров для классификации диалектов, велика вероятность переобучения. Поэтому было принято решение использовать метод случайного исключения активных нейронов (dropout). Итоговая архитектура модели классификации изображена на рисунке 1. Экспериментальным образом были подобраны параметры обучения слоя классификации: скорость обучения - 10^{-6}; количество эпох – 5; вероятность исключения нейрона – 0,5; размер пакета обучения в рамках итерации - 2. Результаты обучения можно увидеть на графиках рисунков 2 и 3.

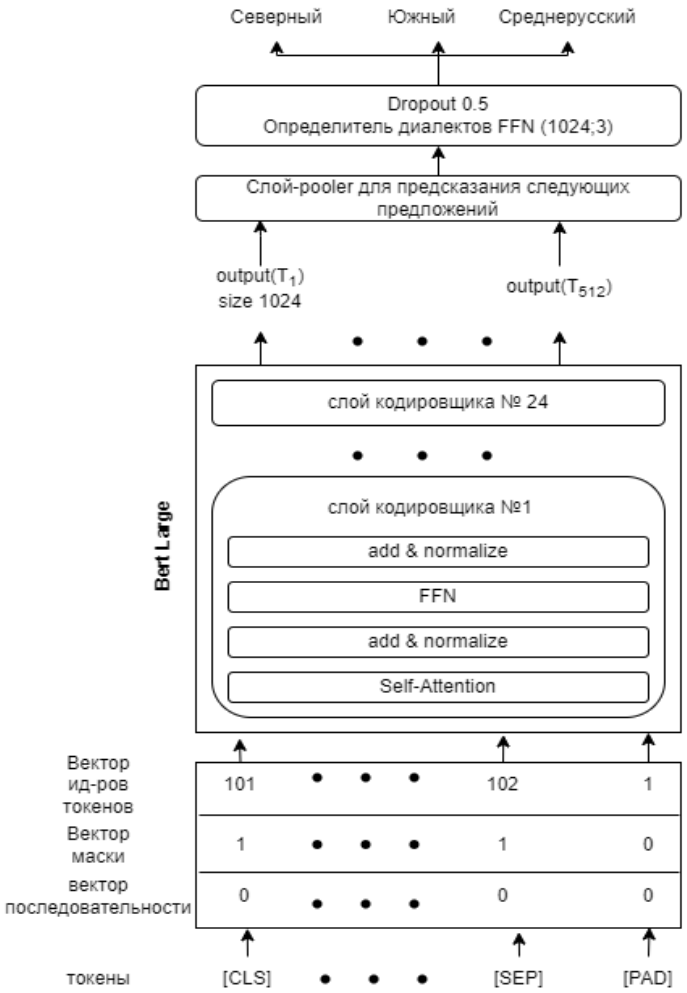

Рисунок 1. Архитектура нейросетевой модели классификации диалектов

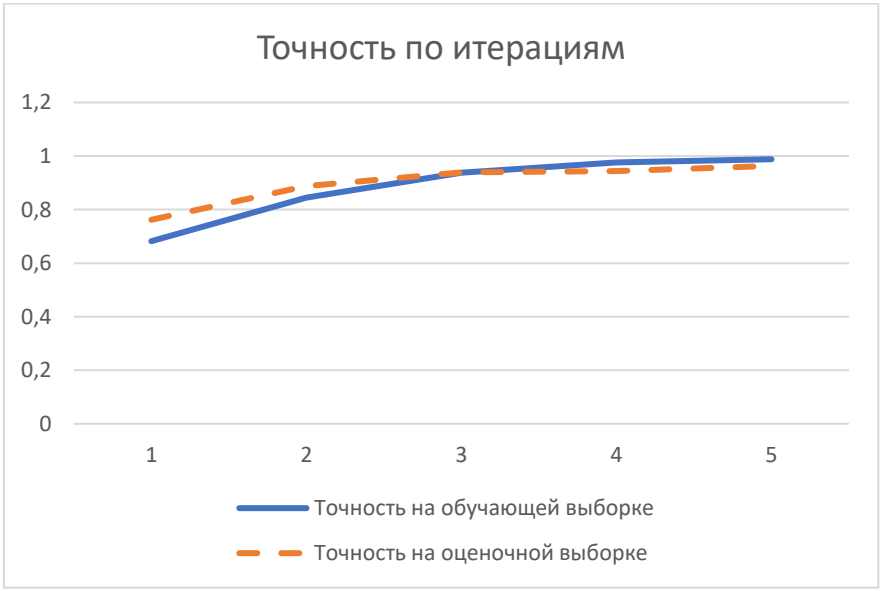

Рисунок 2. Диаграмма точности классификации диалекта во время обучения

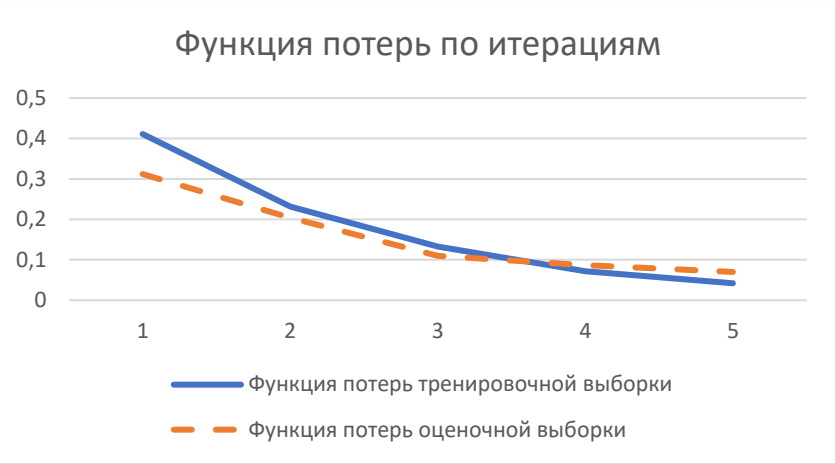

Рисунок 3. Диаграмма функции потерь во время обучения

Определение диалектизмов

Ключевой проблемой обучения нейросетевой модели для поиска диалектизмов является недостаточное количество данных о диалектизмах в контексте. Поэтому было принято решение использовать гибридный подход, основанный на нечёткой логике и метриках модели Bert.

Нечёткая логика позволяет строить системы правил, основанные на нечётких переменных и их функциях принадлежности к множеству. В качестве ответа такой системы выступает не явное решение задачи классификации, а вектор действительных чисел степени принадлежности к тому или иному множеству классифицируемых объектов.

$$F_t(x,a,b,c,d) = \begin{cases} 0, \text{ если } x < a \\ \frac{x-a}{b-a}, \text{если } a \le x \le b \\ 1, \text{ если } b \le x \le c \\ \frac{d-x}{d-c}, \text{если c} \le x \le d \\ 0, \text{ если } x > d \end{cases} \quad (1)$$

Было принято решение использовать кусочно-линейную трапециевидную функцию, так как данная функция имеет простое построение и идеально подходит для прототипирования системы правил. Трапециевидная функция (1) задаётся четырьмя параметрами: a, b, c, d. Параметры функции подбирались экспериментально. Остальные классы функций подходят для более тонкой настройки системы экспертами предметных областей. Их стоит использовать, если известно точное распределение классифицируемых объектов.

Для построения системы правил были сформулированы 4 лингвистические переменные, которые будут далее описаны набором (β, T, X), где β – название лингвистический переменной, T – набор термов или нечётких переменных, X – область допустимых значений (универсальное множество): частота употребления в контексте (β_1); относительная частота появления в диалекте i ($\beta_{2i}\ \ i = \overline{1..3}$); косинусное сходство вложения текста со средним вектором вложения диалекта i ($\beta_{3i}\ \ i = \overline{1..3}$); принадлежность слова к диалекту β_{4i}.

Лингвистическая переменная «Частота употребления в контексте»

Данная лингвистическая переменная будет обозначаться в системе правил как β_1. Смысл данной переменной заключается в том, чтобы описать шанс появления в контексте предложения через предсказание замаскированного слова с помощью модели Bert. Финальный слой нейросети производит моделирование языка, то есть производит классификацию замаскированного слова по токенам словаря. Таким образом, выход нейроной сети обрабатывают функцией softmax, чтобы смоделировать вероятности появления токена.

Тем не менее, использование вероятностей в явном виде приводит к оперированию маленькими действительными числами. Более того нижняя граница вероятностной оценки модели Bert, когда слово считается «не использующимся в контексте» будет зависеть от размера словаря, что крайне важно для функции принадлежности данной нечёткой переменной. Чтобы компенсировать это потребуется провести нормализацию относительно максимального элемента: $normalize(Output)_i = \frac{Output_i - \min(Output)}{\max(Output) - \min(Output)}$.

Однако важно помнить, слово может состоять из нескольких токенов. В таком случае определим частоту встречи слова ($freq_w$) в контексте как среднее гармоническое от всех частот токенов ($freq_{t_i},\ t_i \in T_w, i = \overline{1..n}$) этого слова вычисляется по формуле: $freq_w = \frac{n}{\sum_{i=1}^{n} \frac{1}{freq_{t_i}}}$.

Таким образом, область допустимых значений лексической переменной «Частота употребления в контексте» - $X \in [0; 1]$. Данная область разделяется между следующими термами (нечёткими переменными) с трапециевидными функциями принадлежности: «Не использующееся в контексте» α_{11}, f_t с параметрами: $a = 0,\ b = 0,\ c = 0,\ d = 5 * 10^{-10}$; «Редко использующееся в контексте» α_{12}, f_t с параметрами: $a = 4 * 10^{-10},\ b = 0{,}1,\ c = 0{,}2,\ d = 0{,}35$; «Часто используется в контексте» α_{13}, f_t с параметрами: $a = 0{,}3,\ b = 0{,}8,\ c = 1,\ d = 1$.

Лингвистическая переменная «Относительная частота появления в диалекте»

Лингвистическая переменная «Относительная частота появления в диалекте» считается для каждого диалекта и обозначается в системе правил как $\beta_{2i}, i = \overline{1..3}$. Смысл данной лингвистической переменной описать насколько часто диалектизм встречается в диалекте i относительно общего числа встреч во всех диалектах. Тогда формула вычисления метрики: $relativeFreq_i(w) = \begin{cases} \frac{count_i(w)}{\sum_{k=1}^{3} count_i(W)}, \text{если } \sum_{k=1}^{3} count_i(W) > 0 \\ 0, \text{если } \sum_{k=1}^{3} count_i(W) = 0 \end{cases}$,

где $count_i(w)$ – зафиксированное количество классификаций слова w как диалектизма диалекта i. В случае если у системы нет информации о слове w как о диалектизме: $count_i(w) = 0$. Данная статистика будет собираться при обработке обучающей и тренировочной выборке. Так же система будет давать возможность экспертам внести новые данные для обновления статистики диалектизмов. Всё это поможет экспертам напрямую влиять на результаты работы алгоритма, построенного на нечёткой логике.

Исходя из формулы метрики очевидно, что область допустимых значений лингвистической переменной «Относительная частота появления в диалекте» - $X \in [0; 1]$. Данная переменная будет иметь следующие термы с трапециевидными функциями

принадлежности на своём универсуме: «Не используется в диалекте i» α_{21i}, f_t с параметрами: $a = 0,\ b = 0,\ c = 0,\ d = 0{,}05$; «Используется в диалекте i» α_{22i}, f_t с параметрами: $a = 0,\ b = 0{,}2,\ c = 0{,}3,\ d = 0{,}5$; «Часто используется в диалекте i» α_{23i}, f_t с параметрами: $a = 0{,}4,\ b = 0{,}5,\ c = 1,\ d = 1$.

Лингвистическая переменная «Косинусное сходство вложения текста со средним вектором вложения диалекта»

Лингвистическая переменная «Косинусное сходство вложения текста со средним вектором вложения диалекта» считается для каждого диалекта так же, как и «Относительная частота появления в диалекте». В системе нечётких правил данная переменная будет обозначаться как β_{3i}.

Метрика данной лингвистической переменной считается по формуле косинусного сходства векторов: $simularity(D^i, S) = \frac{D^i * S}{\|D^i\| * \|S\|} = \frac{\sum_{j=1}^{n} D_j^i * S_j}{\sum_{j=1}^{n} \sqrt[2]{D_j^i} * \sum_{j=1}^{n} \sqrt[2]{S_j}}$, где D_i – средний вектор вложений предложений, относящихся к диалекту i; S – вектор вложения предложения, которому принадлежит замаскированное слово. Первым важным примечанием является то, что вложение формируется по исходному предложению до использования процедуры маскирования, чтобы вложение содержало полную информацию о контексте. Второе примечание связано с тем как именно формируется вложение предложения для формирования среднего вложения по диалекту: $D^i = \frac{\sum_{j=1}^{n} S_j^i}{n}$, где S_j^i – вложение j предложения в обучающих данных принадлежащее к i диалекту, а n – общее число предложений обучающей выборки, принадлежащих диалекту i) и участия в вычислении косинусного сходства (S).

Для подсчёта метрики косинусного сходства будет использоваться подход взвешенного среднего. Роль весов играет вектор маски токенов, который является частью входа модели Bert. Таким образом, формула вложения предложения будет иметь вид:

$s_i = \frac{\sum_{j=1}^{512} Output_{ji} * M_j}{\sum_{j=1}^{512} M_j}$, где $i = \overline{1..1024}$; M – вектор масок токенов, состоящий из 0 и 1; Output – матрица выхода модели Bert.

Так как данная метрика ограничена значениями косинуса, область значения лингвистической переменной «Косинусное сходство вложения текста со средним вектором вложения диалекта» - $X \in [0; 1]$. Данная переменная будет иметь следующие термы с трапециевидными функциями принадлежности на своём универсуме: «Отлично от среднего вложения диалекта i» α_{31i}, f_t с параметрами: $a = 0,\ b = 0,\ c = 0,\ d = 0{,}3$; «Незначительное сходство со средним вложением диалекта i» α_{32i}, f_t с параметрами: $a = 0{,}25,\ b = 0{,}4,\ c = 0{,}45,\ d = 0{,}55$; «Схоже со средним вложением диалекта i» α_{33i}, f_t с параметрами: $a = 0{,}5,\ b = 0{,}8,\ c = 1,\ d = 1$.

Лингвистическая переменная «Принадлежность слова к диалектам»

Лингвистическая переменная «Принадлежность к диалектам» обозначается γ. Данная переменная служит логическим выводом системы правил нечёткой логики модели Мамдани. Универсум данной переменной - $X \in [0; 1]$. Универсум разделяется следующими нечёткими переменными: «Не принадлежит к диалектам» λ_1, f_t с параметрами: $a = 0,\ b = 0,\ c = 0{,}2,\ d = 0{,}3$; «Возможно принадлежит к диалектам» λ_2, f_t с параметрами: $a = 0{,}2,$

$b = 0{,}4,\ c = 0{,}5,\ d = 0{,}7$; «Принадлежит к диалектам» λ_3, f_t c параметрами: $a = 0{,}6,\ b = 0{,}8,\ c = 1,\ d = 1$.

Система правил логического вывода для модели Мамдани

Основой логического вывода в нечёткой логике является система правил вывода. Причём правила составления данной системы правил зависит от самого механизма нечёткой логики. В данной работе будет использована модель Мамдани.

Логика построения системы правил базируется на нескольких предположениях:

1) Приоритет имеет переменная «Относительная частота появления в диалекте» (β_{2i}), так как это точная статистическая метрика. Более того, эксперты смогут добавлять в системы новые диалектизмы, тем самым напрямую влияя на статистику слова, а, следовательно, и на логический вывод модели Мамдани.
2) Если слово не встречается в диалекте, то упор идёт на метрики полученные с помощью модели Bert (β_1 и β_{3i}), которые безусловно являются менее надёжными, но позволяют сделать логический вывод в ситуации неизвестности. Причём переменная «Частота употребления в контексте» (β_1) имеет приоритет над переменной «Косинусное сходство вложения текста со средним вектором вложения диалекта» (β_{3i}), так как задача предсказания является первичной для модели Bert.
3) В случае встречи слова, которое предсказывается моделью Bert с крайне низкой вероятностью или токена которого вовсе нет в словаре модели, и при этом отсутствует статистика по встречаемости данного слова в диалектах, и вектор вложения предложения отличается от среднего вектора вложений диалектов, тогда система находится в ситуации полной неизвестности. Было принято решения трактовать подобное слово как потенциальный диалектизм (λ_2).

Согласно предположениям, была сформирована следующая система правил для механизма вывода нечёткой логики:

$$i = \overline{1..3}$$
$$\beta_1 \in \alpha_{11} \cap \beta_{2i} \in \alpha_{21i} \cap \beta_{3i} \in \alpha_{31i} \to \gamma \in \lambda_1$$
$$\beta_1 \in \alpha_{11} \cap \beta_{2i} \in \alpha_{21i} \cap \beta_{3i} \in \alpha_{32i} \to \gamma \in \lambda_2$$
$$\beta_1 \in \alpha_{11} \cap \beta_{2i} \in \alpha_{21i} \cap \beta_{3i} \in \alpha_{33i} \to \gamma \in \lambda_3$$
$$\beta_1 \in \alpha_{11} \cap \beta_{2i} \in \alpha_{22i} \cap \beta_{3i} \in \alpha_{31i} \to \gamma \in \lambda_2$$
$$\beta_1 \in \alpha_{11} \cap \beta_{2i} \in \alpha_{22i} \cap \beta_{3i} \in \alpha_{32i} \to \gamma \in \lambda_3$$
$$\beta_1 \in \alpha_{11} \cap \beta_{2i} \in \alpha_{22i} \cap \beta_{3i} \in \alpha_{33i} \to \gamma \in \lambda_3$$
$$\beta_{2i} \in \alpha_{23i} \to \gamma \in \lambda_3$$
$$\beta_1 \in \alpha_{12} \cap \beta_{2i} \in \alpha_{21i} \to \gamma \in \lambda_1$$
$$\beta_1 \in \alpha_{12} \cap \beta_{2i} \in \alpha_{22i} \to \gamma \in \lambda_3$$
$$\beta_1 \in \alpha_{13} \cap \beta_{2i} \in \alpha_{21i} \to \gamma \in \lambda_1$$
$$\beta_1 \in \alpha_{13} \cap \beta_{2i} \in \alpha_{22i} \to \gamma \in \lambda_4$$
$$\beta_1 \in \alpha_{11} \cap \beta_{21} \in \alpha_{211} \cap \beta_{31} \in \alpha_{311} \cap \beta_{22} \in \alpha_{212} \cap \beta_{32} \in \alpha_{312} \cap \beta_{23} \in \alpha_{213} \cap \beta_{33} \in \alpha_{313} \to \gamma \in \lambda_2$$

Набор данных

Набор данных для обучения и тестирования нейросетевой модели был взят из диалектического корпуса НКРЯ. Доступный набор данных был не сбалансирован (рисунок 4) и состоял из 2142 примерах. Для поиска диалектизмов использовался поиск по тегам

dialex, dialtem, dialsfx и dialform. Первые три записи загруженных данных можно увидеть в таблице 1.

В первоначальных данных нет сведений о диалектах, однако колонка «Title» помимо названия содержит сведения о месте публикации или нахождения текста. Соответственно, с помощью регулярных выражений из колонки «Title» было выделено название субъекта РФ. Далее, используя диалектологическую карту 1965 [Захарова, К. Федоровна, В. Г. Орлова, 2004], можно примерно отнести субъекты РФ по географическому расположению к тому или иному диалекту.

Всего загруженная выборка содержала данные по 7 субъектам РФ, которые были разбиты по 3 диалектам следующим образом:

1) Северный диалект: Вологодская область, Архангельская область;
2) Среднерусский диалект: Псковская область, Пермский край;
3) Южный диалект: Смоленская область, Ставропольский край, Самарская область.

Предобработанные данные были разделены на обучающую, оценочную и тестовую в соотношении 8:1:1. Оценочная выборка была выделена для проверки переобучения во время эпох обучения.

index	Full context	Center	Title
0	развяжут это фсе, повернёш на другу сторону,	другу	О выделке льна (деревня Усть-Поча, Плесецкий р...
1	высохнё,	высохнё	О выделке льна (деревня Усть-Поча, Плесецкий р...
2	потом начьнут в овине или на пецьке, ну в ови...	поставя	О выделке льна (деревня Усть-Поча, Плесецкий р...
3	И мялка была така, на ношках,	така	О выделке льна (деревня Усть-Поча, Плесецкий р...

Таблица 1. Структура данных загруженных из НКРЯ

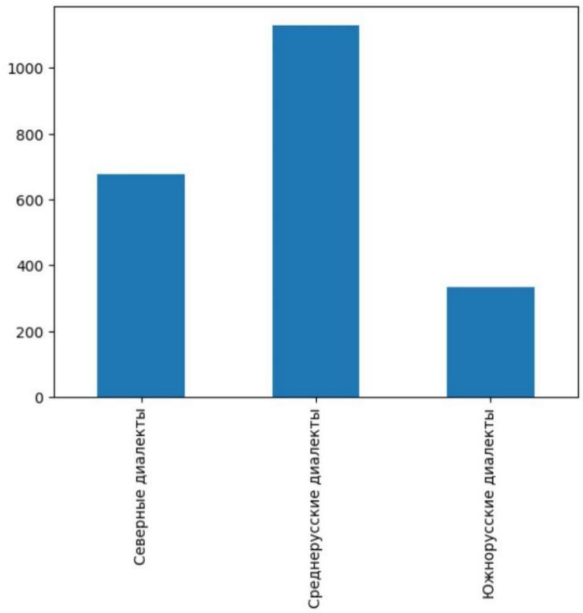

Рисунок 4. Распределение диалектов в обработанном наборе данных.

Оценка результатов

Настроенная на задачу классификации диалектов модель Bert показала точность 95,3% на тестовой выборке, состоящей из 215 примеров. Полученный результат нельзя назвать объективным из-за малого набора данных, что зачастую приводит к переобучению. Для более точной оценки эффективности модели в будущем потребуется расширить тестовый набор данных.

Вторым этапом тестирования является оценка эффективности определения диалектизмов. Для проведения тестирования было принято решения использовать весь набор пред обработанных данных, который был использован для обучения и тестирования тонкой настройки модели Bert под задачу классификации диалектов.

Из каждой записи набора данных (таблица 1) составлялись тестовые случаи. Если текст записи состоял только из диалектизма, то составлялся один позитивный пример. Если был предоставлен полноценный контекст диалектизма, то составлялся позитивный и негативный пример. Так, положительный случай состоит из диалектизма из колонки «Center» и текста («Full context»), негативный - из случайного слова из контекста, исключая диалектизм, и текста. Проводился строгий подсчёт верных классификаций, поэтому класс «Возможно принадлежит к диалектам» считался заведомо не верным. Точность классификации (accuracy) на 4279 тестовых примерах составила 0,4162.

На полученную точность напрямую влияют подобранные параметры функции принадлежности к множеству. Можно попробовать улучшить результат проведя больше экспериментов с ними.

Вторым важным фактором является наличие класса «Возможно принадлежит к диалектам», который ухудшает результат строгой классификации, но помогает в качестве рекомендаций выделять слова из текста.

Заключение

Результатом данной работы является экспериментальная реализация системы классификации диалектов и определения диалектизмов. Созданная система способна классифицировать диалекты русского языка введённого текста и в рекомендательном порядке показывать список найденных диалектизмов.

Экспериментальность данной системы заключается в гибридном подходе к определению диалектизмов, где система правил нечёткой логики опирается на статистические метрики и метрики, полученные с помощью языковых моделей.

Преимуществом предложенной архитектуры системы является адаптируемость под другой язык, в частности, полученные наработки могут быть использованы в системе диалектологического атласа тюркских языков, разрабатываемой в Институте прикладной семиотики АН РТ.

В дальнейшем планируется более точно настроить функции принадлежности термов лингвистических переменных для системы правил нечёткой логики. Для этого потребуется много экспериментов и помощь экспертов в данной сфере. Так же планируется дополнить набор данных для обучения и тестирования.

Использованная литература:

1. A. Etman and A. A. L. Beex, Language and Dialect Identification: A survey, 2015 SAI Intelligent Systems Conference (IntelliSys), London, UK, 2015, pp. 220-231, https://doi.org/10.1109/IntelliSys.2015.7361147
2. Iancu, Ion. A Mamdani type fuzzy logic controller. Fuzzy logic-controls, concepts, theories and applications 15.2 (2012): 325-350.
3. Jauhiainen T, Lindén K, Jauhiainen H. Language model adaptation for language and dialect identification of text. Natural Language Engineering. 2019;25(5):561-583. https://doi.org/10.1017/S135132491900038X
4. McBratney, Alex B., and Adrian W. Moore. Application of fuzzy sets to climatic classification. Agricultural and forest meteorology 35.1-4 (1985): 165-185.
5. Mitchell, Melanie, and David C. Krakauer. The debate over understanding in AI's large language models. Proceedings of the National Academy of Sciences 120.13 (2023): e2215907120.
6. Sun, C., Qiu, X., Xu, Y., Huang, X. (2019). How to Fine-Tune BERT for Text Classification? In: Sun, M., Huang, X., Ji, H., Liu, Z., Liu, Y. (eds) Chinese Computational Linguistics. CCL 2019. Lecture Notes in Computer Science (), vol 11856. Springer, Cham. https://doi.org/10.1007/978-3-030-32381-3_16
7. Zadeh LA. Outline of a new approach to the analysis of complex systems and decision processes. IEEE Trans Syst Man Cybernet 1973, 3:28–44.
8. Zimmermann, H-J. Fuzzy set theory. Wiley interdisciplinary reviews: computational statistics 2.3 (2010): 317-332.
9. Д. Антюхов. Обучение модели естественного языка с BERT, блог компании SberDevices, 2020. Habr. https://habr.com/ru/company/sberdevices/blog/527576/
10. Захарова, Капитолина Федоровна, and Варвара Георгиевна Орлова. Диалектное членение русского языка. УРСС, 2004.