**МЕТОД ДЛЯ ИССЛЕДОВАНИЯ КЛАССИФИКАЦИИ НАУЧНЫХ ТЕКСТОВ
НА РУССКОМ И КАЗАХСКОМ ЯЗЫКАХ****METHOD FOR STUDYING THE CLASSIFICATION OF SCIENTIFIC TEXTS IN
RUSSIAN AND KAZAKH LANGUAGES**

**ИСМУКАНОВА А.Н.¹
КУБИГЕНОВА А.Т.²
ОЖИБАЕВА З.М.²
ЕСМАГАМБЕТОВА Г.К.²**

**ISMUKANOVA A.N.¹
KUBIGENOVA A.T.²
OZHIBAYEVA Z.M.²
ESMAGAMBETOVA G.K.²**

¹*Омский государственный университет
им. Ф.М. Достоевского*

¹*Omsk State University named after F. M.
Dostoyevsky*

²*Кокшетауский государственный
университет им. Ш. Уалиханова*

²*Kokshetau State University named after Sh.
Ualikhanov*

aigera_ismukan@mail.ru

aigera_ismukan@mail.ru

doi: 10.18411/pivsn-29

idsp: pivsn-29

АННОТАЦИЯ

Классификация научных текстов на русском и казахском языках посредством присвоения им универсального десятичного кода (УДК) является актуальной задачей. Задача классификации научных текстов прекрасно решается для текстов, написанных на английском языке в силу простоты морфологии и синтаксиса этого языка. В случае с русским языком имеется ряд нерешенных задач, как видно из нескольких проведенных ранее исследований, которые фокусировались на возможности применять различные подходы. Что же касается казахского языка, то использование аналогичных приемов здесь практически не исследовано.

Ключевые слова: Матрица относительных частот, классификация, кластеризация, граф.

ABSTRACT

The classification of scientific texts in the Russian and Kazakh languages, by means of the assignment of a universal decimal code (UDC), is an issue relevant in the context of international dissemination of Russian and Kazakh research. The problem of the classification of scientific texts is easily solved in the case of the English language due to the simplicity of the latter's morphology and syntax. In the case of the Russian language, some unresolved issues remain, as several investigations of applicability of different approaches were conducted [1-4], whilst the possibility of applying classification methods to texts written in the Kazakh language has not been investigated at all.

Keywords: Matrix of relative frequencies, classification, clustering, count.

ВВЕДЕНИЕ

Классификации современных научных текстов, написанных на русском и казахском языках в XXI веке, необходимо уделить особое внимание в исследованиях. Во многих областях научной деятельности опубликованные тексты являются канонизированным видом продукта научных знаний. Рассмотрение казахского и русского языков в данном контексте обусловлено тем, что языковый союз здесь

является видом языковой интеграции, обусловленной длительным сосуществованием казахского и русского народов в сходных исторических условиях и заключающейся в появлении у этих непосредственных и дальнеродственных языков признаков вторичного сходства [7].

Достижение поставленной цели и решение задач, обозначенных в кандидатской работе, осуществлялись с помощью следующих научных методов: эмпирического исследования (наблюдение и сравнение); теоретического исследования; контент - анализа литературы, интернета и других источников; системного анализа научного контента, в т.ч. системного исследования теоретических разработок отечественных и иностранных ученых в области разработки систем автоматизации.

Одним из перспективных методов, позволяющих получать данные о смысле приведенного русского и казахского текста, является метод латентного семантического анализа (ЛСА), описанный в Листинге 1. Функция токенизации.

В данной работе рассмотрен метод, который в основании переходов между научными текстами, позволяет выявлять группы единиц текстов, имеющих близкую структуру переходов.

ФУНКЦИЯ ТОКЕНИЗАЦИИ В PYTHON

Токенизацией называют разделение данных на мелкие части, т.е. токены. Слова и знаки пунктуации относятся к токенам. Тексты представляются в виде массива нужных слов. Тогда в массивах значимых слов производим чистку на предмет знаков пунктуации и не значимых слов.

Листинг 1. Функция токенизации

```
# -*- coding: utf-8 -*-
import nltk
import string
from nltk.corpus import stopwords

def tokenize_me(file_text):
    #firstly let's apply nltk tokenization
    tokens = nltk.word_tokenize(file_text)

    #let's delete punctuation symbols
    tokens = [i for i in tokens if ( i not in string.punctuation )]

    #deleting stop_words
    stop_words = stopwords.words( 'kazak' )
    stop_words.extend(['не', 'мынау', 'солай', 'онда', 'болу', 'қалай', 'мен', '—', 'оған', 'ал'])
    tokens = [i for i in tokens if ( i not in stop_words )]

    #cleaning words
    tokens = [i.replace("«", "").replace("»", "") for i in tokens]
    return tokens
```

Стандартная функция из Python библиотеки - `nltkword_tokenize()` была использована для работы с многими языками, но нам нужны были только русский и казахский язык.

ЦЕЛЬ ИССЛЕДОВАНИЯ

Целью данного исследования является создание подхода к решению задачи автоматической классификации научных текстов по предоставленным тематическим материалам. Исследованы возможные задачи для классификации научных текстов.

В настоящей работе использованы следующие методы:

- Матрица относительных частот парной встречаемости единиц текста;
- Метод, основанный на индуктивном построении классификаторов для казахского языка;

- Метод, основанный на индуктивном построении классификаторов для русского языка;
- Кластеризация текста с помощью метода ЛСА;
- Метод линейно-дискриминантного анализа.

РЕЗУЛЬТАТЫ И ИХ ОБСУЖДЕНИЕ

Матрица относительных частот

В основе этого метода предполагается создание графа «сильных связей» посредством конструкции матрицы относительных частот, парной встречаемости единиц текста. Для получения такой матрицы необходимо [6]:

- Использовать единицу научного текста, для которой будут конструироваться переходы;
- Последовательность единиц текста перепрограммировать в соответствующий класс;
- Рассчитать частоты парной встречаемости для каждой пары классов с учетом направления развертывания текста.

Далее авторы предполагают, что можно выделить группы текстов, которые имеют близкую структуру. Например, стиль научного текста отличается по определенным грамматическим связям, терминам, которые появляются в тексте чаще других. Соответственно, реже используемые связи между классами единиц текста имеют более низкие частоты встречаемости, поэтому их можно не рассматривать [1, 68].

По полученной матрице строится взвешенный ориентированный граф, структуру которого можно проанализировать, выделив некоторые общие с графами, построенными на основе других текстов или групп текстов.

На основе других научных текстов или групп текстов можно провести анализ по полученной матрице построения взвешенного ориентированного графа.

Обозначим граф, построенный на основании относительной частоты парной встречаемости классов, $G(X, V)$, где X – множество вершин (классов), а V – множество дуг (переходов между классами). Направление дуги зависит от порядка появления элемента в тексте, вес – от относительной частоты встречаемости перехода. «Сильный граф» $G_A(X_1, V_1)$ строится на основе графа $G(X, V)$. В него входят только те дуги, вес которых превосходит назначенный порог α . Очевидно, то чем больше величина порога α , тем меньше вершин и дуг содержит граф сильных связей $G_A(X_1, V_1)$.

Полученные таким образом «сильные графы» рассматриваемых текстов сравниваются. Для этого вводится понятия узла. Узлом называется такая вершина «сильного графа», в которую входит более чем β (заданное число) дуг. Далее используется коэффициент Роджерса-Танимато:

$$\rho_{1,2} = n_{\downarrow 1,2} / ((n_{\downarrow 1} + n_{\downarrow 2} - n_{\downarrow 1,2}))$$

Где n_1, n_2 – число узлов в 1-м и 2-м сравниваемых текстах соответственно, а $n_{1,2}$ – число общих узлов в сравниваемых текстах.

Значение коэффициента ρ заключено в пределах от 0 до 1. В том случае, если сравниваемые тексты не имеют общих узлов, ρ равно 0, а если множества их узлов совпадают, ρ равно 1. Чем больше доля общих узлов, тем ближе значение к 1. Близость «сильных графов» означает близость текстов в некотором смысле.

Отличия между матрицами переходов. Метод подсчета отличий между матрицами переходов может быть использован для выделения групп текстов. Приведем его описание. Сначала определяется матрица относительных частот парной встречаемости классов единиц текста, как описано выше в пункте 2.1 «Метод «сильного графа». Затем вводится индекс различия P , который вычисляется для пары

текстов или групп текстов как сумма мер различия по всем соответствующим элементам двух матриц:

$$P = \sum_{i=1}^n \sum_{j=1}^n |a_{i,j} - b_{i,j}|$$

Где $A = |a_{ij}|$, $i = 1..n$, $j = 1..n$, $B = |b_{ij}|$, $i = 1..n$, $j = 1..n$ матрицы относительных частот парной встречаемости единиц двух любых текстов, n – число выбранных для анализа классов. Получая близкие значения индекса различия P , может сделать вывод о близости текстов в некотором смысле.

Информация, вычисляемая на основании выборок размером в 300-400 слов: рассмотрение этой характеристики может выделить тексты, которые менее сложны для восприятия, в частности за счет преимущественного использования более коротких слов.

Будем использовать для определения длины слова в буквах выборки в 500 слов, считая, что такая выборка несет достаточно большую информацию о колебаниях языковых частот, при чем в этом случае из рассмотрения исключаются тексты, которые имеют размер меньше чем 500 слов. По всем выборкам получим среднее значение для выборочного среднего и выборочной дисперсии.

Для каждого текста получаем:

- n – число выборок;
- $\bar{m}$ – средняя длина слова в выборке;
- s^2 – выборочная дисперсия;
- s – выборочное среднее квадратичное отклонение.

Аналогичные расчеты проводим для контрольной группы текстов, с которой происходит сравнение.

Будем считать, что выборки из контрольной группы текстов и тестируемых текстов достаточно однообразны, получены из нормального распределения, поэтому проведем тест исключительности с помощью критерия Стьюдента, проверив, являются ли статистически значимыми различия в средних полученных выборок и выборок, полученной на основе контрольной группы. Рассмотрим критерий в виде:

$$t = \frac{\bar{m}_1 - \bar{m}_2}{sd} \sqrt{\frac{n_1 * n_2}{n_1 + n_2}}$$

Где $\bar{m}_1$ и $\bar{m}_2$ – сравниваемые средние величины, n_1 и n_2 число выборок в двух сравниваемых текстах, индекс 1 для текста, который рассматривается, а 2 для текстов контрольной группы, а sd – несмещенная оценка среднего квадратичного отклонения в двух сравниваемых текстах, вычисляемая для сравнения двух средних частот по формуле:

$$sd = \sqrt{\frac{(n_1 - 1) * s_1^2 + (n_2 - 1) * s_2^2}{n_1 + n_2 - 2}}$$

Можно исключить возможность того, что рассматриваемый научный текст близок к контрольной группе, если найденная величина превышает критические значения для распределения Стьюдента при заданном уровне значимости. Научные тексты, которые имеют значение меньше, чем самое большое значение критерия для текста из контрольной группы, определяем максимальным коэффициентом, полученным для текста из контрольной группы, и исключаем все тексты, для которых этот уровень выше.

Общее распределение длины слова.

Рассмотрим общее распределение длины слов для каждого текста, сравним с данными, полученными для текстов из контрольной группы.

Вначале получим информацию о распределении длины слов во всех научных текстах, то есть сколько в каждом тексте слов, имеющих по паре и более букв. Нужно рассматривать те словоформы, которые встречаются в тексте, без приведения к начальной форме. В группу слов включим все слова, которые имеют большую длину.

Ставим вопрос, какова вероятность того, что распределения длин слов в буквах у текста взяты из той же «генеральной совокупности», что и у контрольной группы, то есть могут рассматриваться как управляемые одними и теми же закономерностями.

Используем непараметрический критерий Колмогорова-Смирнова, измеряющий разницу между накопленными относительными частотами в сравниваемых текстах по формуле:

$$\lambda = |d \max| \sqrt{\frac{n_1 * n_2}{n_1 + n_2}}$$

где $|d \max|$ обозначает максимальную разницу между накопленными относительными частотами, n_1 и n_2 – количество рассматриваемых слов в сопоставляемых текстах, индекс 1 для текста, который рассматривается, а 2 для текста контрольной группы.

ЗАКЛЮЧЕНИЕ

В работе рассмотрены алгоритмы, описанные в предыдущих пунктах, исключая возможность близости текстов, для которых получено значение k к текстам из контрольной группы.

Модель ЛСА, несмотря на трудоемкость и непрозрачность, может использоваться для разного ряда задач при обобщении или расширении смысла поискового запроса.

Способы классификации научных текстов на русском и казахском языках и присвоение им универсального десятичного кода (УДК) достигаются несколькими математическими методами. Согласно названию, в статье описываются разные методы для определения термина и слов. Преимущества этих методов описаны кратко, они только очерчены.

БЛАГОДАРНОСТИ

Выражаем глубокую признательность кандидату технических наук Лаврову Дмитрию Николаевичу за обсуждение статьи. Замечания, им высказанные, позволили существенно улучшить изложение полученных результатов.

1. Ньюман Д., А. Асунсьон, П. Смит и М. Веллинг. 2007. Распределенный вывод для скрытого распределения Дирихле. Достижения в системах обработки нейронной информации 20: 17-24. , стр. 68.
2. Розова С.С. Классификационная проблема в современной науке. Новосибирск: Наука. Сиб. отд-ние, 1986 г., с. 199-200.
3. Medhat W., Hassan A., Korashy H. Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journ. 2014, no. 5, pp. 1093–1113.
4. Polyakov I.V., Sokolova T.V., Chepovsky A.A., Chepovsky A.M. Text classification problem and features set. Vestn. NGU. Ser.: Informatsionnye tekhnologii [Novosibirsk State Univ. Journ. of Information Technologies]. 2015, vol. 13, iss. 2, pp. 55–63 (in Russ.).
5. Интернет пещур: <https://psibook.com/linguistics/sravnitelnyy-analiz-effektivnosti-algoritmov-raspoznavaniya-avtorstva-tekstov-po-chastotam-perehodov.html>
6. Зеленцов Б.П. Матричные модели функционирования оборудования систем связи //Вестник СибГУТИ.2015.№ 4.
7. Словарь лингвистических терминов: Изд. 5-е, испр-е и дополн. — Назрань: Изд-во "Пилигрим". Т.В. Жеребило. 2010.