УДК 007:681.512.2

БИНАРНЫЕ ИЕРАРХИЧЕСКИЕ ЧИСЛА ДЛЯ ВЫЧИСЛЕНИЯ СЕМАНТИЧЕСКОЙ БЛИЗОСТИ ПРЕДЛОЖЕНИЙ ЕСТЕСТВЕННОГО ЯЗЫКА

И. Ю. Каширин, д.т.н., профессор кафедры ВПМ РГРТУ Рязань, Россия;
orcid.org/0000-0003-1694-7410, e-mail: igor-kashirin@mail.ru

Рассматривается новая технология вычисления семантической близости предложений естественного языка, предварительно обработанных обученными нейронными сетями. Для программной реализации семантического анализа используется инструментарий Spacy и WordNet.

В качестве предметной области выбрана автоматическая верификация новостных материалов политической тематики.

Для вычисления числовых параметров семантической близости используется теория бинарных иерархических чисел. Приведены основные операции с иерархическими числами. Рассмотрен принцип минимизации сложных семантических отношений таксономии. Иерархические числа используются при анализе родовидовой таксономии предметной области естественно-языкового предложения.

Экспериментальная часть исследований проведена для тестового программного обеспечения, реализованного на языке Python v.3 (Anaconda 3). В качестве исходных текстов новостных статей использованы материалы международных изданий WSJ, PBS News Hour, AC News и других.

Выполненная серия экспериментов дает возможность оценить рассматриваемую технологию как технологию вычисления семантической близости предложений, не уступающую по эффективности имеющимся современным международным аналогам.

***Целью работы** является создание новой технологии, применяемой при автоматизированном вычислении семантической близости конструкций естественного языка для формирования тематических подборок электронных новостных материалов.*

***Ключевые слова**: бинарные иерархические числа, семантическая близость, родовидовая таксономия, интеллектуальная обработка данных, база знаний, семантические сети, анализ естественного языка, нейронные сети.*

DOI: 10.21667/1995-4565-2023-86-110-121

Введение

Автоматический анализ семантики текстов [1], в том числе в таких сложных предметных областях, как электронные новостные ресурсы [2,3], становится весьма эффективным программным инструментарием при решении задач интеллектуальной обработки информации. Подавляющее большинство высокотехнологических решений в этой области основано на лексико-семантической базе данных WordNet [4], содержащей семантические сети для различных предметных областей. Инструментарий WordNet, включающий API-средства [5], использует несколько онтологических таксономий, позволяющих, в частности, оценивать семантическую близость понятий, заданных словоформами естественного языка.

Применение алгоритмов вычисления семантической близости понятий и предложений используется в весьма разнообразных областях искусственного интеллекта, в частности это такие проблемы, как:

- классификация документов [6];
- решение задач лексической сабституции [7];
- автоматическое разрешение лексико-семантической омонимии [8];
- перефразирование предложений [9];
- аннотирование или упрощение текстов [10];
- разработка интеллектуальных агентов информационного поиска [11];

– проектирование естественно-языковых справочных систем [12];
– анализ тональности текста [13];
– интеллектуальная идентификация фейковых новостей [3].

В то же время при всем многообразии существующих методов [14] вычисления семантической близости предложений проблема получения новых эффективных алгоритмов остается актуальной, поскольку современные методы зачастую не дают достоверных результатов.

Задачей настоящей статьи является описание новой технологии вычисления семантического сходства в предложениях естественного языка. Технология основана на применении теории иерархических чисел, основы которой впервые опубликованы в [15].

В качестве предметной области для исследования технологии выбрана сфера автоматизации обнаружения фейковых новостей [16,17], являющаяся также весьма актуальной для современных информационных технологий.

Теоретическая часть

Автоматический анализ предложений естественного языка

Архитектуры современных автоматизированных систем, использующих нейронные сети для анализа естественно-языковых текстов, как правило, не отличаются разнообразием. Примерный вид такой архитектуры приведен на рисунке 1.

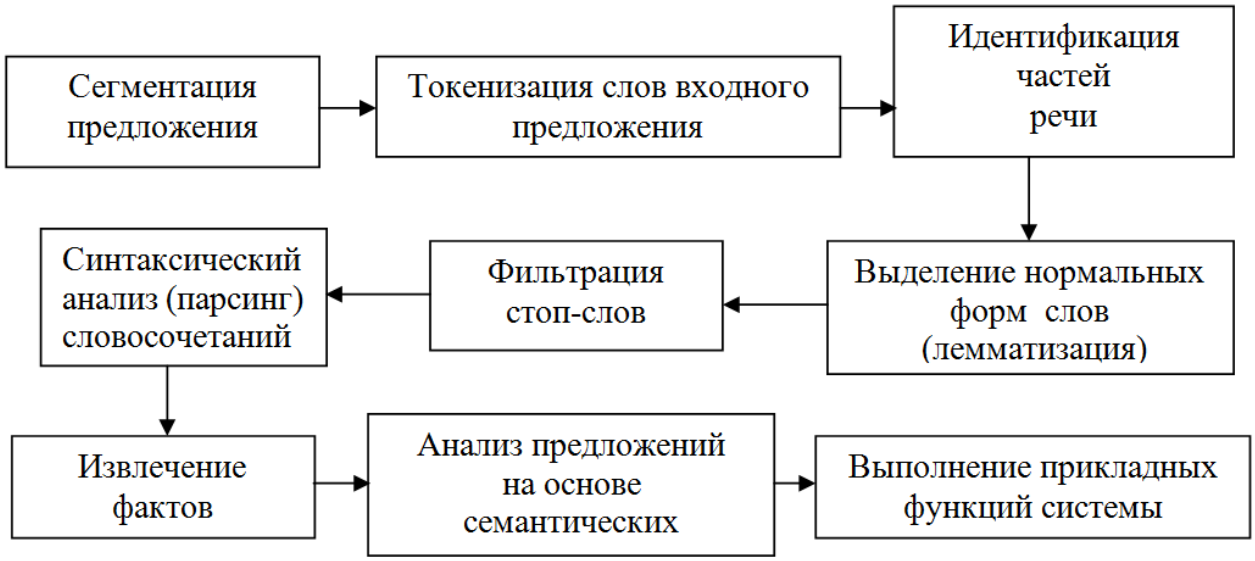


Рисунок 1 – Архитектура систем анализа естественно-языковых текстов
Figure 1 – Architecture of natural language text analysis systems

В приведенной архитектуре первым этапом анализа является сегментация предложения, представляющая собой выделение предложений из окружающего контекста с удалением пунктуации (при необходимости) и побочных конструкций.

На этапе токенизации в предложении выделяются отдельные словоформы для последующего морфологического анализа.

Идентификация частей речи заключается в использовании словаря, а также грамматических и морфологических таблиц аффиксов и окончаний. На этом этапе словоформам приписываются тэги словарных категорий, таких как «существительное», «прилагательное», «наречие» и т.п. Такие тэги принято называть POS («Part-of-Speech») [18].

Для лемматизации как определения нормальных форм слов используются тезаурусы или лексико-семантические базы данных и знаний, например англоязычная лексическая база данных WordNet [5] или русскоязычный тезаурус RuWordNet [19], который можно использовать как семантическую сеть.

Этап фильтрации стоп-слов предназначен для удаления словоформ, включенных в отдельный словарь как неинформативные слова.

Синтаксический анализ выполняет выделение словосочетаний и других локальных конструкций и в некоторых версиях анализаторов опускается. Наиболее известными русскоязычными синтаксическими анализаторами, формирующими выходные конструкции систем составляющих и деревьев зависимостей, являются Natasha [20] и Spacy 3 [21].

Извлечение фактов производится с помощью специализированных образцов (темплейтов) на основе анализа словосочетаний. Во многих реализациях этот этап или является заключительным, или опускается.

Анализ предложений на основе таксономий представляет собой этап, весьма близкий к семантическому анализу (анализу смысла). Упомянутые инструментальные средства WordNet и RuWordNet содержат родовидовые отношения, таксономии которых позволяют вычислять гипероним и гипоним синсета («synset (synonym set)»), соответственно представляющие собой понятие-предок и понятие-потомок. Синсет – это значение конкретного слова, которое включает в себя само слово, а также определение слова и множество его синонимов. Семантическая сеть RuWordNet содержит кроме этого более сложные отношения «Часть-целое» и «Домен-атрибут».

Заключительный этап анализа предложения на естественном языке представляет собой выполнение основных функций прикладной системы, т.е. функций, соответствующих конечной цели анализа, например перевод с одного языка на другой, формирование ответа на вопрос в вопросно-ответных системах [12] или анализ семантического сходства предложений.

В настоящей статье рассматривается технология вычисления семантического сходства предложений, суть этой технологии заключается в следующей последовательности шагов.

1. Каждое предложение разбивается на список токенов.
2. Определяются части речи (POS), часто с учетом контекста.
3. Выбирается наиболее подходящее значение для каждого слова в предложении (устранение омонимии и неоднозначности слов).
4. Вычисляется сходство предложений на основе сходства пар (или троек) слов.

Первые два этапа были описаны ранее, два других реализуются в различных системах разными методами. Рассмотрим некоторые наиболее эффективные из них.

Проблематика вычисления схожести предложений естественного языка

Рассмотрим сначала устранение неоднозначности слов и их смыслового сходства (WSD, Word Sense Disambiguation).

Будем анализировать предложения естественного языка из предметной области «верификация достоверности новостей» [17]. В качестве новостной тематики выберем конфликт на Ближнем Востоке. Тематическая подборка сообщений известных изданий сведена в таблицу 1. Рассмотрим первую новость, представленную рисунком 2.

В этом примере есть словосочетание «не готова», которое может иметь, по крайней мере, три значения в разных контекстах:

- пицца оказалась не готова;
- студенческая группа оказалась не готова к контрольной работе;
- команда оказалась не готова к такому повороту событий.

Что же имеется в виду в сообщении: армия не находится в состоянии боевой готовности, и ей еще нужно готовиться и готовиться, или же армия не была осведомлена о готовящемся нападении?

The Israel Defense Forces were not ready to repel a strike on the border.
Армия обороны Израиля оказалась не готова к отражению удара на границе.

Рисунок 2 – Пример сообщения новостной ленты
Figure 2 – Example of a news feed message

Известен метод WSD с оригинальным *алгоритмом Майкла Леска*, позволяющий устранить эту неоднозначность. Алгоритм основан на множестве определений различных слов – глоссарии. Глоссарий содержит и примеры устойчивых отношений между словами (словосочетания).

Таблица 1 – Тематическая подборка сообщений известных изданий
Table 1 – Thematic selection of messages from well-known publications

Текст новости на языке оригинала	**Русский перевод**	**Новостной ресурс**
The Israel Defense Forces were not ready to repel a strike on the border	Армия обороны Израиля оказалась не готова к отражению удара на границе	*Telegram «Военная хроника»*
The Israeli intelligence services, including the Mossad, were not aware of the impending attack on the country by the Palestinian radical movement Hamas	Разведывательные службы Израиля, в том числе «Моссад» не были осведомлены о готовящемся нападении на страну со стороны палестинского радикального движения ХАМАС	*New York Times*
The weekend's assault, which caught Israel off guard on a major Jewish holiday	Нападение, произошедшее на выходных, которое застало Израиль врасплох	*PBS News Hour*
Facts and testimony suggest that the Netanyahu government knew in advance about the actions of Hamas, which led to the deaths of hundreds of Israelis and Palestinians	Факты и свидетельские показания позволяют предположить, что правительство Нетаньяху заранее знало о действиях ХАМАС, которые привели к гибели сотен израильтян и палестинцев	*AC News*
Israel knew about the Gaza attack in advance, top congressman says.	По словам высокопоставленного конгрессмена, Израиль заранее знал о нападении в Газе.	*Palestine Chronicle*
The problems with the Palestinian Authority are related to gas, a field discovered offshore in front of the Gaza Strip	Проблемы с Палестинской автономией связаны с газом, месторождением, открытым на шельфе перед сектором Газа	*Neftegaz.ru*
WSJ reported on the preparation of 2 thousand US military personnel for deployment in the Middle East	WSJ сообщила о подготовке 2 тысяч военных США к развертыванию на Ближнем Востоке	*WSJ*

Для определения смыслового сходства нужно сравнить количество одинаковых слов в разных определениях: чем больше число общих слов, тем ближе эти слова по смыслу.

Для приведенного примера алгоритм определяет, что словосочетания «не готова» и «к отражению» часто встречаются вместе, выделяя этим контекстом смысл «не осведомлена».

Однако этот алгоритм проверяет попарно *все* слова предложения и оказывается вычислительно сложным, а также не всегда правильно вычисляющим вложенный автором смысл. Кроме того, б*о*льшая часть программной реализации касается части речи «существительные».

Более новой является модификация этого метода, предложенная *Дао и Симпсоном*, использующая измерения *Херста-Сен-Онжа* [22]. Этот алгоритм использует не только синсет-глоссарий, но и значения родственных слов вообще.

При вычислении взаимосвязи между двумя синсетами s1 и s2 пара hype-hype означает, что синсет для гиперонима s1 сравнивается с сисетом для гиперонима s2. Пара hype-hypo означает, что значение для гиперонима s1 сравнивается со значением для гипонима s2:

OverallScore(s1, s2)= Score(hype(s1)-hypo(s2)) +
+ Score(gloss(s1)-hypo(s2)) + Score(hype(s1)-gloss(s2))...
(OverallScore(s1, s2) is also equivalent to OverallScore(s2, s1)).

Например, словосочетание «не готова» имеет три возможных смысла, а «отражение» два смысла. Итогом являются 6 смысловых значений. В алгоритме измеряются перекрытия определений слов с учетом закона Ципфа, частным случаем ленинского принципа экономии мышления: «длина слов обратно пропорциональна их использованию». Затем алгоритм для определения смысла выбирает определение с наибольшим перекрытием определений.

Множество подходов определения семантической близости с помощью нейронных сетей были уже проанализированы в литературе [23,24]. В частности, сравнивались следующие подходы сообщества *DKPro* с инструментарием *DKPro Similarity:*

- n-граммы, использующие меры включения;
- вычисление евклидового расстояния;
- близость на основе коэффициента Жаккара;
- конусное сходство;
- вычисление минимального количества подстановок для унификации двух текстов;
- вычисление количества подстановок с нормализацией;
- покрытие максимального числа токенов непересекающимися цепочками трех слов текстов;
- использование метрики семантического пространства двух текстов.

Бинарные иерархические числа

Кратко идея бинарных иерархических чисел выглядит так. Пусть B – множество чисел с элементами {0, 1}, n ϵ B (n = 0 или n = 1), пусть также есть выделенный символ «.».

Множество A = B ∪ «.» ∪ ʌ определяется как алфавит с целыми числами из B, где «∪» – операция объединения множеств, а ʌ – пустой символ. Тогда грамматика:

$\hat{h} \rightarrow ʌ$, $\hat{h} \rightarrow h$,

$h \rightarrow < n >$, $h \rightarrow < n > . <h>$

описывает множество бинарных иерархических чисел $\hat{H}$ с элементами $\hat{h}$.

Пример бинарного иерархического числа: 0.1.0.1 .

Бинарные иерархические числа представляют собой численные индексы вершин двух двоичных деревьев: положительного и отрицательного с одной общей вершиной 0.

Порождение вершины влево от 0 производится бинарной операцией 0+0 = 0.0, порождение вершины вправо – бинарной операцией 0+1 = 0.1.

Порождение отрицательных вершин выполняется операцией «-» соответственно:

0-0 = -0.0, 0-1 = -0.1.

Графически это выглядит так (рисунок 3):

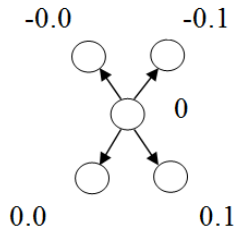


Рисунок 3 – Дерево, соответствующее бинарным иерархическим числам
Figure 3 – Is a tree corresponding to binary hierarchical numbers

Операция порождения трассы «+» может быть более сложной:

0.1 + 1.1 = 0.1.1.1

0 + 0.1 = 0.0.1

Операция, обратная порождению, удаление терминальной вершины «--», является унарной:

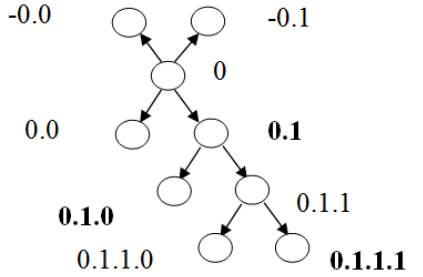
0.1.1.1-- = 0.1.1

0.1.0-- = 0.1

Абсолютное число отсчитывается с начальной вершины 0, относительное число – от любой вершины бинарного дерева. Считать ли число абсолютным или относительным индексом вершины бинарного дерева, определяет человек, решающий прикладную задачу с помощью иерархических чисел. Абсолютный индекс всегда начинается с символа 0.

Можно рассматривать только положительную часть алгебраической системы бинарных иерархических чисел. В этом случае операции, претендующие на получение отрицательного индекса, будут иметь результатом 0.

Вычисление наиболее общей вершины (генерализация, обобщение) интерпретируется как поиск их общего предка (рисунок 4):



0.1.1.1 º 0.1.0 = 0.1.0 º 0.1.1.1 = 0.1

Рисунок 4 – Операция выделения общего предка
Figure 4 – Common Ancestor allocation operation

Умножение положительного числа на отрицательное всегда равно 0.

Понятие обратного элемента (числа) можно описать так. Что можно сказать о числах 0.1.1.0 и 1.0.0.1 ? Они получены полной инверсией атомарных элементов 0 и 1 во всех разрядах. Умножение таких чисел всегда даст нулевой результат. Очевидно, эти числа нельзя считать абсолютными (отсчитываемыми от корня дерева). Однако пути доступа к терминальным вершинам для этих чисел полностью противоположны: если считать «1» положительным выбором, а «0» отрицательным, то всякий спуск в дереве на один уровень ниже для первого числа будет противоположным выбору, определяемому вторым числом.

Одним из существенных метапонятий для бинарных иерархических чисел является *структурный путь* SP между двумя любыми вершинами бинарного дерева, а следовательно, и между двумя соответствующими иерархическими числами. Структурный путь для двух иерархических чисел – это цепочка вершин, которые необходимо пройти, чтобы попасть из одной вершины, соответствующей первому числу, в другую, соответствующую второму. Такой путь может быть не единственным. В качестве примера рассмотрим числа: x.0.0.0 и x.0.0.1.1. Общая вершина-предок, через которую вначале будет проходить подъем в бинарном дереве, а затем начнется спуск в направлении терминальных вершин, вычисляется так:

x.0.0.0 º x.0.0.1.1= x.0.0.

Структурным для них может быть следующий путь:

SP = x.0.0.0 → x.0.0 → x.0.0.1 → x.0.0.1.1 .

Сложность пути PC вычисляется как суммарное количество промежуточных вершин таксономии, лежащих на кратчайшем пути из какой-либо вершины, соответствующей некоторому понятию, X в вершину Y.

Например, для цепочки x.0.0.0→x.0.0→x.0.0.1→x.0.0.1.1сложность пути

PC (x.0.0.0→x.0.0→x.0.0.1→x.0.0.1.1) = PC(SP) = 4.

Поскольку путей из одной вершины в другую может быть более чем один, перебором сравнений может быть найден кратчайший структурный путь для любых двух бинарных

иерархических чисел. Правило минимизации, описывающее вычисление минимального пути, выглядит так:

$<x,y \in T , PC(x \rightarrow y) = \mu> \Rightarrow \ <PC(x \rightarrow y) = \tau>, \quad \{\exists \mu, \mu < \tau\},$

где T – множество бинарных иерархических чисел в заданной таксономии.

Характеристика сложности пути PC для двух понятий есть числовая оценка их смысловой близости.

Семантическая близость двух понятий, например, в родовидовой таксономии определяется длиной пути наследования, сложность которого вычисляется разностью бинарных иерархических чисел. При проектировании таксономий из нескольких возможных структур нужно выбирать более простую структуру. Этот принцип описывается следующим выражением:

$$\left| \bigcup_{i=1}^{m} \bigcup_{j=n-1}^{1} Is - A(x_i^j, x_i^{j+1}) \right| = \tau \quad \Rightarrow$$

$$\left| \bigcup_{i=1}^{m} \bigcup_{j=k-1}^{1} Is - A(x_i^j, x_i^{j+1}) \right| = \mu \quad [\exists \mu, \mu < \tau],$$

где m – суммарное количество терминальных вершин таксономии, x_i^1 – корневая вершина, x_i^n – терминальные вершины, μ и τ – меры семантического сходства двух концептов.

Технология вычисления сходства предложений на основе бинарных иерархических чисел

В качестве нового подхода к вычислению семантической близости предложений и словосочетаний естественного языка предлагается использовать определенное выше правило минимизации.

Опишем новую технологию последовательностью шагов ее реализации.

Предлагается в качестве примера использовать предметную область верификации политических новостей. Для получения естественно-языковых новостных текстов на первом шаге технологии используется инструментарий newscatcher 0.2.0 для среды Python 3.0 [25].

```
!pip install newscatcher
from newscatcher import Newscatcher
news_source = Newscatcher('nytimes.com')
last_news_list = news_source.news
article = last_news_list[0]
article.keys()
dict_keys(['title', 'title_detail', 'links', 'link', 'id', 'guidislink',
           'summary','summary_detail', 'published', 'published_parsed',
           'tags', 'media_content','media_credit', 'credit'])
print(article.title)
print(article.summary)
print(article.link)
https://www.nytimes.com/2020/03/24/world/coronavirus-updates-maps.html
print(article.published)
```

Таким способом можно не только выбрать новостные статьи, но и создать корпус [22] для обучения нейронных сетей, формирующих тезаурусы и семантические сети, аналогичные WordNet [22] и RuWordNet [19], рассмотренные ранее. В нашем случае для анализа были выбраны предложения, сведенные в таблицу 1. В настоящей статье пока не ставится целью верификация выбранных новостей. Вычисление семантического сходства здесь необходимо *для выделения новостных материалов одной тематики*, которые в будущем можно будет исследовать на достоверность. Вторым и третьим шагами технологии являяются токенизация и лемматизация предложений, выделенных из новостной ленты.

```
import nltk
from nltk.tokenize import word_tokenize
from nltk.stem import WordNetLemmatizer
nltk.download('averaged_perceptron_tagger')
from nltk.corpus import wordnet
sentence = " Israel knew about the Gaza attack in advance,
               top congressman says. "
def pos_tagger(nltk_tag):
 if nltk_tag.startswith('J'):
     return wordnet.ADJ
 elif nltk_tag.startswith('V'):
     return wordnet.VERB
 elif nltk_tag.startswith('N'):
     return wordnet.NOUN
 elif nltk_tag.startswith('R'):
     return wordnet.ADV
 else:
     return None
pos_tagged = nltk.pos_tag(nltk.word_tokenize(sentence))
wordnet_tagged = list(map(lambda x: (x[0], pos_tagger(x[1])), pos_tagged))
lemmatized_sentence = []
 for word, tag in wordnet_tagged:
  if tag is None:
     lemmatized_sentence.append(word)
 else:
     lemmatized_sentence.append(lemmatizer.lemmatize(word, tag))
lemmatized_sentence = " ".join(lemmatized_sentence)
```

На четвертом шаге технологии можно произвести разметку слов предложений в форме деревьев зависимостей с применением инструментария, аналогичного Natasha [20] и Spacy 3 [21]. Подробное применение этого инструментария приведено в [17].

Пятый шаг технологии связан с использованием родовидовой таксономии для политических новостей, фрагмент которой приведен на рисунке 5.

На рисунке иерархические числа соответствуют цифровым индексам синсетов для существительных в родовидовой таксономии абстрактного уровня. В конкретных новостных материалах при достаточно полном тезаурусе появятся подстановки конкретизации: σ1={«Страна» ⇒«Израиль»; «Столица» ⇒ «Иерусалим»; «Руководитель» ⇒ «Нетаньяху»; «Народ» ⇒ «Израильтяне»; «Армия⇒ «Цахал»; «Разведка⇒ «Моссад»}; σ2={«Страна» ⇒ «Палестина»; «Столица» ⇒ «Газа»; «Руководитель» ⇒ «Аббас»; «Народ» ⇒ «Палестинцы»; «Армия» ⇒ «Хамас»; «Разведка» ⇒ NIL}.

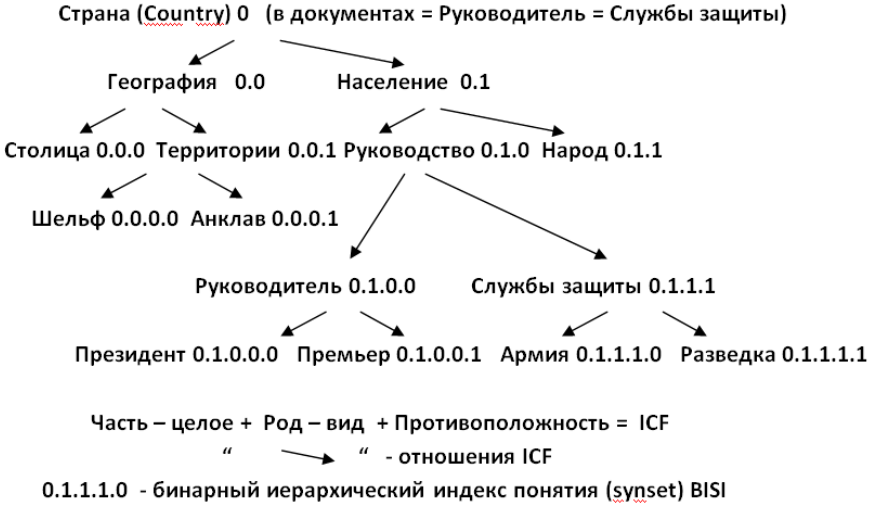


Рисунок 5 – Фрагмент таксономии для анализа политических новостей
Figure 5 – Fragment of taxonomy for analyzing political news

Бинарные числа для каждого конкретизированного экземпляра будут полностью соответствовать числам абстрактной таксономии рисунка 5. Т.е., например, («Моссад», 0.1.1.1.0), («Палестинцы», 0.1.1). Сложность пути РС, например, для отношения «Армия»-«Народ» {«Армия»→«Народ», РС(0.1.1.1.0→0.1.1.1→0.1.0→0.1→0.1.1)=5}, а сложность пути «Страна»-«Руководитель»{«Страна»→«Руководитель», РС(0→0.1→0.1.0→0.1.0.0)=4}. Второй путь является менее сложным, следовательно, в предложениях словоформу «Нетаньяху» можно заменить словоформой «Израиль» с большей вероятностью, чем «Моссад» заменить на «Израильтяне». В этом суть ранее рассмотренного принципа минимизации.

Идея нового метода сопоставления предложений по их смыслу состоит в переборе различных вариантов подстановок словоформ новостного текста словоформами абстрактной и конкретной таксономий с учетом правила минимизации. Вычисление семантической близости В ($\sum word_i$) можно производить с помощью известного инструментария, например WordNet, с полным перебором возможных подстановок из таксономии. Однако вычисленные значения при этом умножаются на коэффициент $(1 / \sum PC(word_i)+1)*B$. Здесь $\sum word_i$ – сумма величин семантического сходства всех слов предложения, вычисленных инструментарием WordNet, а $\sum PC(word_i)$ – сумма всех значений РС слов, участвовавших в подстановке.

Экспериментальные исследования

Эксперименты производились с предложениями, представленными в таблице 1. Квадратные матрицы попарного сходства предложений в sklearn из Python и технологии РС, представленные далее, показывают повышение характеристик сходства с учетом аналогичного понимания человеком на 5-18 % в пользу новой технологии.

```
 sentencesWordNet=['Netanyahu government knew in advance about the
 actions of Hamas, which led to the deaths of hundreds of
 Israelis and Palestinians',
 'Israel knew about the Gaza attack in advance, top congressman says',
'The Israel Defense Forces were not ready to repel a strike on the border']
vectorizer = CountVectorizer()
sentence_vectors = vectorizer.fit_transform(sentencesWordNet)
similarity_matrix = cosine_similarity(sentence_vectors, sentence_vectors)
print(similarity_matrix)
[[1.         0.33593551 0.23973165]
 [0.33593551 1.         0.23354968]
 [0.23973165 0.23354968 1.        ]]
sentencesPC = ['Israel knew in advance about the actions of Hamas, which led
  to the deaths of hundreds of Israelis and Palestinians',
  'Israel knew about the Gaza attack in advance, top congressman says',
  'The Israel was not ready to repel a strike on the border']
vectorizer = CountVectorizer()
sentence_vectors = vectorizer.fit_transform(sentencesPC)
similarity_matrix = cosine_similarity(sentence_vectors, sentence_vectors)
print(similarity_matrix)
[[1.         0.39886202 0.31448545]
 [0.39886202 1.         0.2508726 ]
 [0.31448545 0.2508726  1.        ]]
```

Прикладное значение полученных результатов заключается в получении новой технологии вычисления семантического сходства естественно-языковых предложений. Новая технология является более эффективной в сравнении с ранее известными зарубежными аналогами.

Библиографический список

1. **Astrakhantsev N.A., Turdakov D.Yu.** (2013), Automatic construction and enrichment of informal ontologies: A survey. Programming and Computer Software 39(1): 34–42.

2. **Каширин И.Ю.** Нейронные сети для идентификации пользователя на основе анализа посещений новостного сайта // Вестник Рязанского государственного радиотехнического университета. 2022. № 82. С.104-111. DOI: 10.21667/1995-4565-2022-82-104-111.

3. **Каширин И.Ю.** Идентификация достоверности новостей с помощью моделей машинного обучения // Вестник Рязанского радиотехнического университета. 2023. № 83. С.36-47. DOI: 10.21667/1995-4565-2023-83-36-47.

4. **Artese M.T.** Ensemble-Based Short Text Similarity: An Easy Approach for Multilingual Datasets Using Transformers and WordNet in Real-World Scenarios. September 2023 Big Data and Cognitive Computing 7(4):158.

5. WordNet. A Lexical Database for English. [Electronic resource]. Update date:20th September, 2023 URL: https://wordnet.princeton.edu/download/current-version (date of application: 17.10.2023).

6. **Shahul ES**. Document Classification: 7 Pragmatic Approaches for Small Datasets. [Electronic resource]. Update date: 4th September, 2023 URL: https://neptune.ai/blog/document-classification-small-datasets (date of application: 17.10.2023).

7. **Леонтьев А.В. , Митрофанова О.А.** Разработка и оценка алгоритма лексической субституции для русского языка на основе предсказывающих нейросетевых моделей // Terra Linguistica. 2023, Том 14, № 2. С. 31–44.

8. **Кобрицов Б.П., Ляшевская О.Н., Шеманаева О.Ю.** Снятие лексико-семантической омонимии в новостных и газетно-журнальных текстах: поверхностные фильтры и статистическая оценка. [Электронный ресурс]. 2005. Дата обновления: 7.03.2005. URL: https://elar.urfu.ru/bitstream/10995/1388/1/IMAT_2005_03.pdf (дата обращения: 17.10.2023).

9. **Xianggen L. , Mou L., Fandong M., Hao Z., Song J.** (2020). "Unsupervised Paraphrasing by Simulated Annealing". Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics: 302–312.

10. **Cao, Z.**, et al. (2015). Learning summary prior representation for extractive summarization. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). Vol. 2. (pp. 829–833).

11. **Albrecht S., Stone.P**. (2018). Autonomous Agents Modelling Other Agents: A Comprehensive Survey and Open Problems. Artificial Intelligence, Vol. 258, pp. 66-95.

12. **Kashirin I.Yu., Khoroshevsky V.F.** Development of an ATN-oriented linguistic processor. Suwalki,15-21 June, Poland, 1987г.

13. **Lin Ya., Liang Ch., Xu J., Yang Ch.,Wang Yo.** 2022. Zhixiaobao at semeval-2022 task 10: Apporoaching structured sentiment with graph parsing. // Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pp.1343–1348.

14. The Complete Guide to Text Similarity with Python – NewsCatcher . [Electronic resource]. Update date: 5 September, 2022 URL: https://www.newscatcherapi.com/blog. (date of application: 17.10.2023).

15. **Каширин И.Ю.** Иерархические числа для проектирования ICF-таксономий искусственного интеллекта // Вестник Рязанского радиотехнического университета. 2020. № 71. С.71-82. DOI: 10.21667/1995-4565-2020-71-71-82.

16. **Van Der Linden S., Roozenbeek J., Compton J.** 2020. Inoculating against fake news about 855 covid-19. Frontiers in psychology, p. 2928.

17. **Каширин И.Ю.** Индикаторы семантических признаков в моделях Data Minig для вычисления фейк ранга новостей VI международный научно-технический форум СТНО-2023. сборник трудов // Международный научно-технический форум СТНО-2023. Сборник трудов. Том 4. С.15-19.

18. Part-of-Speech Tagging [Electronic resource]. Update date: 12.10.2022 URL: https://melaniewalsh.github.io/Intro-Cultural-Analytics/05-Text-Analysis/13-POS-Keywords.html. (date of application: 17.10.2023).

19. **Loukachevitch N., Lashevich G.** Multiword expressions in Russian Thesauri RuThes and RuWordNet. Proceedings of the AINL FRUCT 2016, 2016. pp.66-71.

20. Natasha — качественное компактное решение для извлечения именованных сущностей из новостных статей на русском языке. [Электронный ресурс]. 2023. Дата обновления: 09.02.2023. URL: https://natasha.github.io/ner/ (дата обращения: 09.02.2022).

21. Можно всё: решение NLP задач при помощи spacy. [Электронный ресурс]. 2023. Дата обновления: 09.02.2023. URL: https://habr.com/ru/post/531940/ (дата обращения: 09.02.2022).

22. WordNet-based semantic similarity measurement. [Электронный ресурс]. 2023. Дата обновления: 8.10.2022. URL: https://www.codeproject.com/Articles/11835/WordNet-based-semantic-similarity-measurement (дата обращения: 18.10.2023).

23. **Крюкова А.В.** Определение семантической близости текстов с использованием инструмента DKPro Similarity // Компьютерная лингвистика и вычислительные онтологии. ITMO University. 2018. ITMO University. С.87-97.

24. **Сатыбалдина Д.Ж., Овечкин Г.В., Калымова К.А**. Система распознавания статических жестов рук с использованием камеры глубины // Вестник Рязанского государственного радиотехнического университета. 2020. № 72. С. 93-105.

25. 4 Python Web Scraping Libraries To Mining News Data – NewsCatcher. [Electronic resource]. Update date: 28.11.022 URL: https://www.newscatcherapi.com/blog/python-web-scraping-libraries-to-mine-news-data. (date of application: 19.10.2023).

UDC 007.681.512.2

BINARY HIERARCHICAL NUMBERS TO CALCULATE SEMANTIC PROXIMITY OF NATURAL LANGUAGE SENTENCES

I. Yu. Kashirin, Dr. in technical sciences, full professor, department of computational and applied mathematics, RSREU, Ryazan, Russia;
orcid.org/0000-0003-1694-7410, e-mail: igor-kashirin@mail.ru

The article discusses a new technology for calculating semantic proximity of natural language sentences preprocessed by trained neural networks. For software implementation of semantic analysis, Spacy and WordNet tools are used. Automatic verification of political news materials was chosen as subject area.

The theory of binary hierarchical numbers is used to calculate numerical parameters of semantic proximity. Basic operations with hierarchical numbers are given. The principle of minimizing the taxonomy complex semantic relations is considered. Hierarchical numbers are used when analyzing generic taxonomy of subject area of natural language sentence. The experimental part of the research was carried out for test software implemented in Python v.3 (Anaconda 3). Source texts of news articles made use of the materials from international publications such as WSJ, PBS News Hour, AC News and others.

The performed series of experiments makes it possible to evaluate the technology in question as a technology for calculating semantic proximity of sentences, which is not inferior in efficiency to existing modern international analogues.

***The aim** of the work is to create a new technology used in automated calculation of semantic proximity of natural language constructions for the formation of thematic collections of electronic news materials.*

***Keywords**: binary hierarchical numbers, semantic proximity, generic taxonomy, intelligent data processing, knowledge base, semantic networks, natural language analysis, neural networks.*

DOI: 10.21667/1995-4565-2023-86-110-121

1. **Astrakhantsev N.A., Turdakov D. Yu.** (2013), Automatic construction and enrichment of informal ontologies: A survey. *Programming and Computer Software*. 39(1): 34-42.

2. **Kashirin I.Yu.** Nejronnye seti dlya identifikacii pol'zovatelya na osnove analiza poseshchenij novostnogo sajta. *Vestnik Ryazanskogo gosudarstvennogo radiotekhnicheskogo universiteta.* 2022, no. 82, pp.104-111. (in Russian). DOI: 10.21667/1995-4565-2022-82-104-111.

3.**Kashirin I.Yu.** Identifikaciya dostovernosti novostej s pomoshch'yu modelej mashinnogo obucheniya. *Vestnik Ryazanskogo gosudarstvennogo radiotekhnicheskogo universiteta.* 2023, no 83, pp.36-47. (in Russian). DOI: 10. 21667/1995-4565-2023-83-36-47.

4. **Artese M.T.** Ensemble-Based Short Text Similarity: An Easy Approach for Multilingual Datasets Using Transformers and WordNet in Real-World Scenarios. September 2023 *Big Data and Cognitive Computing* 7(4):158.

5. *WordNet. A Lexical Database for English.* [Electronic resource]. Update date:20th September, 2023 URL: https://wordnet.princeton.edu/download/current-version (date of application: 17.10.2023).

6. **Shahul ES**. *Document Classification: 7 Pragmatic Approaches for Small Datasets*. [Electronic resource]. Update date: 4th September, 2023 URL: https://neptune.ai/blog/document-classification-small-datasets (date of application: 17.10.2023).

7. **Leont'ev A.V., Mitrofanova O.A.** Razrabotka i ocenka algoritma leksicheskoj sub-stitucii dlya russkogo yazyka na osnove predskazyvayushchih nejrosetevyh modelej. *Terra Linguistica*. 2023, vol. 14, no. 2, pp. 31-44.

8. **Kobricov B.P., Lyashevskaya O.N., Shemanaeva O.Yu.** Snyatie leksiko-semanticheskoj omonimii v novostnyh i gazetno-zhurnal'nyh tekstah: poverhnostnye fil'try i statisticheskaya ocenka. [Electronic resource]. 2005. Update date: 7.03.2005. URL: https://elar.urfu.ru/bitstream/10995/1388/1/IMAT_2005_03.pdf (date of application: 17.10.2023).

9. **Xianggen L., Mou L., Fandong M., Hao Z., Song J.** (2020). «Unsupervised Paraphrasing by Simulated Annealing». *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics: 302-312.

10. **Cao, Z.** et al. (2015). Learning summary prior representation for extractive summarization. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing* (Volume 2: Short Papers). Vol. 2. (pp. 829-833).

11. **Albrecht S., Stone.P**. (2018). Autonomous Agents Modelling Other Agents: A Comprehensive Survey and Open Problems. *Artificial Intelligence*. Vol. 258, pp. 66-95.

12. **Kashirin I.Yu., Khoroshevsky V.F.** *Development of an ATN-oriented linguistic processor*. Suwalki, 15-21 June, Poland, 1987.

13. **Lin Ya., Liang Ch., Xu J., Yang Ch.,Wang Yo.** 2022. Zhixiaobao at semeval-2022 task 10: Aporoaching structured sentiment with graph parsing. *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pp.1343-1348.

14. The Complete Guide to Text Similarity with Python – NewsCatcher. [Electronic resource]. Update date: 5 September, 2022 URL: https://www.newscatcherapi.com/blog. (date of application: 17.10.2023).

15. **Kashirin I.Yu.** Ierarhicheskie chisla dlya proektirovaniya ICF-taksonomij iskusstvennogo intellekta. *Vestnik Ryazanskogo gosudarstvennogo radiotekhnicheskogo universiteta*. 2020, no. 71, pp.71-82. (in Russian). DOI: 10.21667/1995-4565-2020-71-71-82

16. **Van Der Linden S., Roozenbeek J., Compton J.** 2020. Inoculating against fake news about 855 covid-19. *Frontiers in psychology*, p. 2928.

17. **Kashirin I.Yu.** Indicators of semantic features in Data Minig models for calculating fake news rank VI International Scientific and Technical Forum STNO-2023. *Proceedings International Scientific and Technical Forum STNO-2023.* Collection of works. vol. 4. pp.15-19.

18. Part-of-Speech Tagging [Electronic resource]. Update date: 12.10.2022 URL: https://melaniewalsh.github.io/Intro-Cultural-Analytics/05-Text-Analysis/13-POS-Keywords.html. (date of application: 17.10.2023).

19. **Loukachevitch N., Lashevich G.** Multiword expressions in Russian Thesauri RuThes and RuWordNet. Proceedings of the AINL FRUCT 2016, 2016. pp.66-71.

20.*Natasha – kachestvennoe kompaktnoe reshenie dlya izvlecheniya imenovannyh sushchnostej iz novostnyh statej na russkom yazyke*. [Elektronnyj resurs]. 2023. Data obnovleniya: 09.02.2023. URL: https://natasha.github.io/ner/ (data obrashcheniya: 09.02.2022).

21. *Mozhno vsyo: reshenie NLP zadach pri pomoshchi spacy*. [Electronic resource]. 2023. Update date: 09.02.2023. URL: https://habr.com/ru/post/531940/ (date of application: 09.02.2022).

22. WordNet-based semantic similarity measurement. [Electronic resource]. 2023. Update date: 8.10.2022 URL: https://www.codeproject.com/Articles/11835/WordNet-based-semantic-similarity-measurement (date of application: 18.10.2023).

23. **Kryukova A.V.** Opredelenie semanticheskoj blizosti tekstov s ispol'zovaniem in-strumenta DKPro Similarity. *Komp'yuternaya lingvistika i vychislitel'nye ontologii. ITMO University*. 2018. ITMO University, pp .87-97.

24. **Satybaldina D.Zh., Ovechkin G.V., Kalymova K.A.** Sistema raspoznavaniya staticheskih zhestov ruk s ispol'zovaniem kamery glubiny. *Vestnik Ryazanskogo gosudarstvennogo radiotekhnicheskogo universiteta*. 2020, no.72, pp. 93-105. (in Russian).

25. *4 Python Web Scraping Libraries To Mining News Data – NewsCatcher*. [Electronic resource]. Update date: 28.11.022 URL: https://www.newscatcherapi.com/blog/python-web-scraping-libraries-to-mine-news-data. (date of application: 19.10.2023).