Инженерный вестник Дона, №4 (2016)
ivdon.ru/ru/magazine/archive/n4y2016/3832

О методе определения текстовой близости основанном на семантических классах

С. Х. Г. Бермудес, С. У. Керимова

Южный федеральный университет, Ростов-на-Дону

Аннотация: Рассматривается метод сравнения текстов на русском языке в подзадаче определения семантической близости. Проводится обзор существующих методов сравнения. Предложен метод определения степени подобия между текстовыми пассажами в пределах семантического класса. Существующие методы сравниваются с предлагаемым методом и сравнением, сделанным людьми, в эксперименте, который показывает адекватность предложенного метода.
Ключевые слова: текстовая близость, семантические классы, сравнение текстов, представление семантических схем, текстовые пассажи.

Введение

В настоящее время поиск сходства между текстами имеет большое практическое применение, в том числе академического обнаружения плагиата и дистанционного образования. В работе [1] упоминаются три основные категории обнаружения текстуального сходства: сравнение на основе слов, линейный поиск на основе пунктов, использующийся поисковыми системами и стилистический анализ. Также существуют методы, основанные на различных характеристиках текстов, такие как методы, основанные на семантике, как для обнаружения плагиата [2-4], так и для поиска информации [5].

Предложенная работа представляет собой подробное описание шага 4 о сравнении на близость общей схемы, представленной в работе [6]. Это промежуточная стадия между схемами семантического представления и оценки текстуального сходства.

На входе процесса сравнения текстов есть два документа, предназначенные для сравнения, один из которых является эталоном. На первом уровне анализа проводится извлечение текстовых пассажей, как это описано в работе [7], выходом этого первого уровня будет перечень значимых пассажей из каждого документа, которые послужат в качестве

входных данных для следующего уровня для разрешения анафоры [6, 8], а последний, в свою очередь, поступает на уровень семантического представления [6, 5] схем. Построенная схема представления является входом для обнаружения уровня семантического сходства. На этом шаге сходство между документами определяется с использованием семантического критерия сравнения подобия, на основе семантических классов слов и заданным эталоном. Общая схема сравнения на близость представлена на рис. 1.

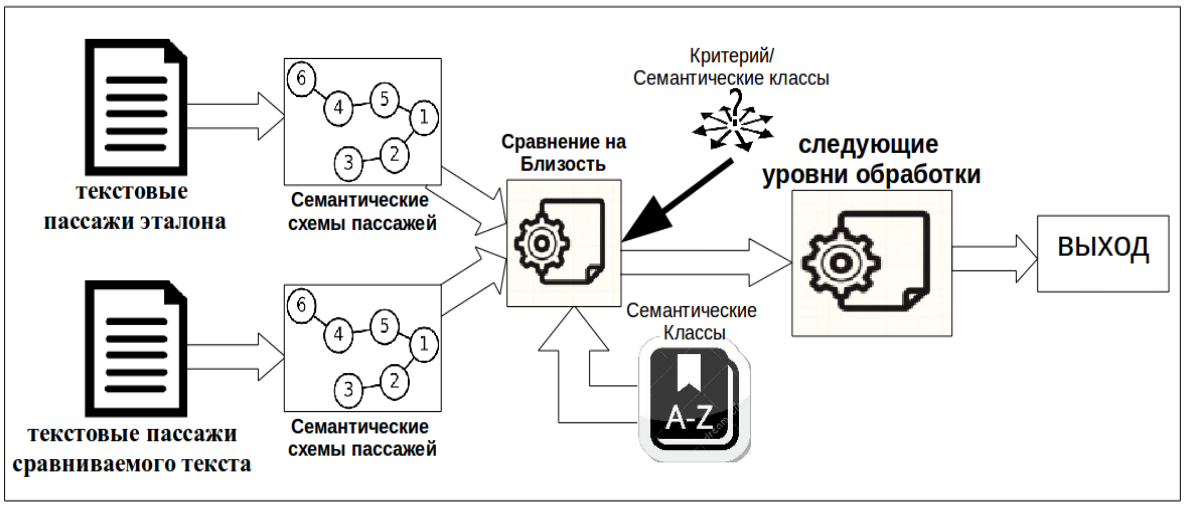

Рис. 1. – Общая схема сравнения на близость

Это означает, чтокритерий семантического сходства текстового эталона и текст для сравнения могут быть представлены через элементы смысла [5] из текстовых пассажей в соответствии с семантическим классом слов, участвующим в сравнении.

Обзор существующих методов и моделей определения сходства текстов

Поиск степени семантической схожести текстов был рассмотрен в качестве задачи в рамках многих международных конференций [9-12]. Многие из разработанных моделей главным образом использовали эмфазу в поисках характеристик, которые совпадают в обоих текстах, обеспечивая тем самым обнаружение того, имеют ли два текста аналогичный смысл.

В работе [13], предложена модель Маркова для определения количества информации, которую разделяют два суждения. В рамках этого исследования разработан набор правил, которые стремятся сделать вывод, если два текста имеют тот же смысл.

Исследование [14], свидетельствует о том, что для того, чтобы измерить степень текстового семантического сходства между этими текстами, они должны быть представлены не терминами, выраженными одинаковыми словами, а терминами, выраженными разными словами. Тогда найти сходство в этом типе представления помогают автоматические переводы, для которых используется инструмент PanLex, который позволяет создать статистический словарь. Если перевод возможен, это означает, что термин эквивалентен термину в тексте, выраженному другими словами.

Другой способ подойти к этой задаче - это рассмотреть ее как проблему QuestionAnswering, где один из текстов является вопросом, а другой ответом. Это суть работы [15], где предлагается модель, которая измеряет степень сходства в функции, если ответ действительно отвечает на вопрос.

Особо следует упомянуть процедуру, указанную в работе [5], где возникает сравнение семантического сходства фрагментов текста и строится семантический критерий сравнения, учитывающий структуру семантических схем, в этом смысле автор объясняет, что: "*Пусть s_q и s_t семантические схемы фрагментов текстов q и t соответственно. Тогда критерий близости φ данных семантических схем определим следующим образом:*"

$$\begin{cases} \varphi(S_q,\mathrm{S}_t) = (S_q \approx S_t); \\ \varphi(S_q,\mathrm{S}_t) \in \mathrm{D}; \\ \mathrm{D}=[0..1]. \end{cases}$$

где: символ ≈ обозначает операцию установления близости, аD - множество значений критерия близости. Если $\varphi\ (s_q,\ s_t) = 1$, то имеет место полная близость, если $\varphi\ (s_q,\ s_t) = 0$, близость отсутствует.

В большинстве задач в момент обработки текстов выполняется некий тип текстового сравнения, в котором слова сравниваются с другими словами, и/или предложения с другими предложениями. Критерии текстового сравнения в работе [5]:

Базовые критерии сравнения близости которые считают частоту встречаемости слов в тексте, сравнивая относительно эталона (запроса).

$$\Phi_{\text{база}} = \frac{p}{q}$$

где p — число совпадающих слов в запросе и фрагменте текста, q - число слов в запросе. Считается, что два слова одинаковы, если их начальные формы совпадают.

Семантические критерии, которые сравнивают предложения и не только считают частоту слов в тексте, сравнивая относительно эталона (запроса), а также рассматривают отношения между фразами, участвующими в сравнении. Например, семантический критерий сравнения на близость:

$$\Phi_{\text{семантик}} = \frac{m}{n}$$

где m – число совпадающих элементов смысла в запросе и фрагменте текста, n – общее число элементов смысла в запросе.

В целом, подходы, описанные в выше, имеют характеристики, которые позволяют выделить три группы. Первая из них считает частоту встречаемости n-грамм символов, слов и некоторых лексических отношений, таких как синонимы и гиперонимы. Кроме того, многие из этих подходов подчеркивают представление естественного языка, чтобы затем использовать алгоритмы сходства между строками, такими как коэффициент подобия Жаккара, который вычисляет количество уникальных терминов совместно

используемых между двумя текстами; косинусного подобия, который измеряет угол между векторами обеих коллекций слов в тексте; и расстояние Левенштейна, которое состоит из минимального количества необходимых операций для трансформации одной цепочки характеристик в другую.

Вторая группа характеристик рассмотрена в упомянутых работах, а также в работе [16], это меры подобия слов, предлагаемых инструментом NLTK на языке программирования Python.В этом случае определяется семантическое сходство между двумя текстами как максимальное значение полученное между парами слов.

Третья группа рассматривает меры на основе Corpus, с использованием показателей, предлагаемых текстовому семантическому сходству [17]. Использование взаимной информации (PMI) для вычисления подобия между парами слов, и латентно-семантического анализа (ЛСА).

Предлагаемый метод определения текстовой близости

Мы будем основываться на формуле для вычисления семантического сходства, предложенной в [5] с помощью семантических классов WordNet(русскоязычный); для выполнения семантического сравнения между семантическими схемами текстовых пассажей.

Особенностью метода является то, что это фрагменты текстов являются текстовыми пассажами с семантическим содержанием [7], а не любые фрагменты, как в работе [5].

Отличием предлагаемого критерия семантической близости текстовых пассажей эталона и сравниваемого текста является вычисление доли совпадающих элементов смысла, в соответствии с семантическим классом слов, участвующих в сравнении.

$$\Phi_{\text{семантик/класс}} = \frac{\sum_i^k \frac{\sum p_j}{l}}{n} \qquad (1)$$

где p – фактор совпадения между словами, участвующих в сравнении, для каждого элемента смысла, согласно семантическому классу в интервале [0,1], p = 1, если слово идентично, p = 0 если слово вне семантического класса и p = (0,1) в зависимости от степени синонимии; l — количество слов каждого элемента смысла; k — количество элементов смысла в текстовом пассаже сравниваемого текста, n — общее число элементов смысла в текстовом пассаже эталона. Необходимо, чтобы эксперт предварительно определял степень синонимии каждого семантического класса. Это может быть сделано по предопределению в эталоне.

На уровне представления текстовых пассажей в семантических схемах получается число n-схем пассажей эталона и число m-схем пассажей сравниваемого текста, которые будут сравнены в соотношении n к m, но совпадения будут считаться в суммарном количестве n, независимо от количества схем сравниваемого текста, таким образом, что если одна схема имеет совпадения с более, чем одной схемой другого текста, это будет считаться главным фактором совпадения.

Для каждой итерации сравнения рассматриваются следующие предварительные условия:

- Каждая смысловая схема текстовых пассажей, как из текста-образца, так и из сравниваемого текста, имеет характеристики и они получены в соответствии с процессом, который объясняется в [5], пример таких схем представлен на рисунке 2, что соответствует фразе «*Но при всем множестве понятий и определений системы, не сформировано понятие метасистемы*».

- Так, для семантической схемы рисунке 2 ее ярусный состав имеет вид:

• яруснулевогоуровняобразуют $\{S(a),S(b),S(c),S(d),S(e),S(f),S(g), S(h)\}$;

• ярус первого уровня образуют вершины $\{r_1, r_2, r_3\}$;

• ярус второго уровня образует вершина $\{r_4, r_5, r_6\}$;

• ярус третьего уровня образует вершина $\{r_7\}$.

• ярус четвёртого уровня образует вершина {r_8}.

• ярус пятого уровня образует вершина {r_9}.

• ярус шестого уровня образует вершина {r_{10}}.

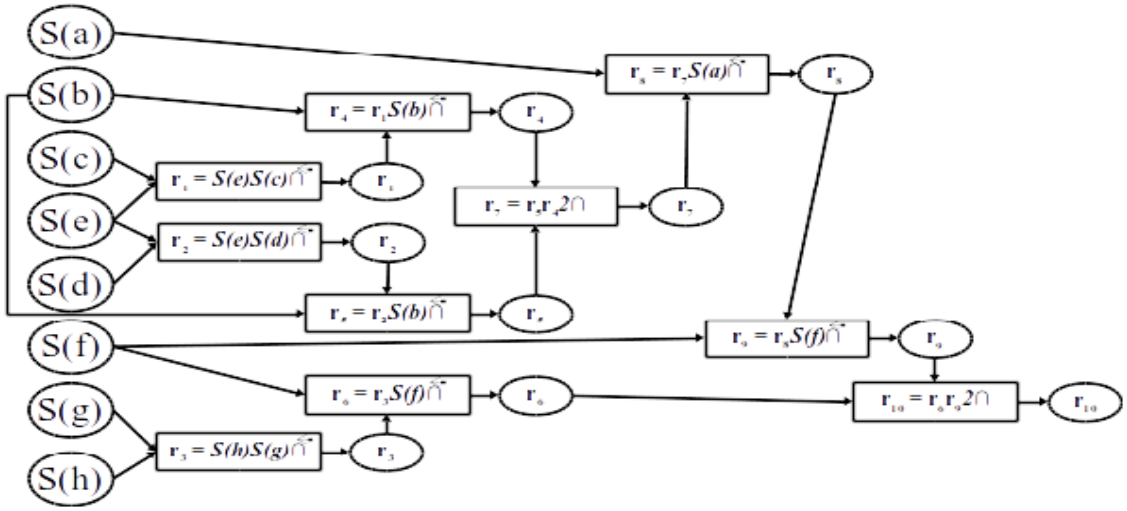

Рис. 2. – Пример семантической схемы текстового пассажа

- Каждое слово текстового пассажа из текста-образца связана со списком слов, которые принадлежат к семантическому классу, согласно WordNet для русского языка, со степенью подобия, назначенной экспертом. Примером может быть, слово «*мировоззрение*», связанными словами и значениями для согласования коэффициента будут: «*миропонимание*: 0,9», «*миросозерцание*: 0,9», «*воззрение*: 0,7», «*мнение*: 0,7»; и «*принципы*: 0,5».

- Для каждого элемента значения семантических схем текстовых пассажей, участвующих в сравнении, фактор совпадения "p" получается из простых средних значений совпадения слов в элементе семантической схемы сравниваемого текста относительно слов из текста-образца.

- Степень сходства между текстовыми пассажами будет определена по формуле (1) в части 3 настоящей статьи.

Пример процесса сравнения представлен ниже, а также его отличие от метода, используемого [5]:

Рассмотрим пассажи эталона и их соответствие семантической схеме, данной на рисунке 2 и пассажи сравниваемого текста: «*Но существование*

большого количества определений системы не привело к формированию понятия метасистемы.», чья схема представлена на рисунке 3.

Таким образом совпадения для метода [5] будут $r_2 \approx r_2 = 1$ у $r_3 \approx r_3 = 1$, применяя формулу $\Phi_{семантик} = 2/10 = 0,2$; то есть 20% сходства.

В то время для предложенного семантического метода классов совпадения: $r_2 \approx r_2 = 1$; $r_3 \approx r_3 = 1$; $r_5 \approx r_4 = 0,83$; $r_6 \approx r_7 = 0,93$; $r_7 \approx r_6 = 0,83$; $r_8 \approx r_8 = 0,58$; $r_9 \approx r_9 = 0,5$ у $r_{10} \approx r_{10} = 0,7$; используя формулу $\Phi_{семантик/класс} = 6,67/10 = 0,667$; то есть 67% сходства.

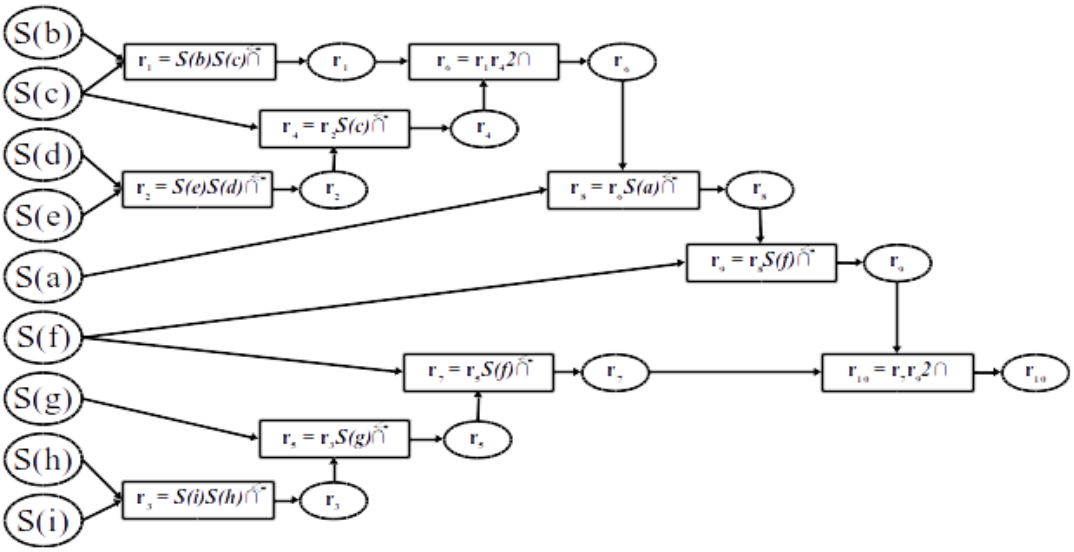

Рис. 3. – Семантическая схема текстового пассажа сравниваемого текста

Сравнение методов

Проведем сравнение некоторых из рассмотренных выше методов с предлагаемым в данной работе методом и анализом, сделанным экспертами. В частности, сравниваются следующие методы и программы:

1. Методы сравнения текстов, основанные на коэффициенте подобия Джакарда, косинусное подобие и расстояние Левенштейна, используя для этого онлайн-программу алгоритмов сходства между цепочками текста, основанных на языке программирования php [18].

2. Метод латентно-семантического анализа и другие методы поиска информации, используя для этого онлайн-программу обнаружения

“plagiarisma.net”; которая основана на использовании поисковых систем “Google”, “Babylon” и “Yahoo”.

3. Программа обнаружении плагиата ЮФУ, которая называется «Анти-плагиат», которая предполагаемо основана на методе поиска анализа скрытой семантики и на других собственных алгоритмах, принадлежащих разработчикам программного обеспечения.

4. Метод определения подобия для поиска информации, который указан в работе [5], который называется «*Ф семантик*».

Для проведения эксперимента были выбраны четыре текста: 1) введение научной статьи под названием «Подход к определению метасистемы как системы» [19]; 2) Модифицированный плагиат из текста один, который был написан специально, путем замены в оригинальном тексте некоторых слов на синонимы и фразы, схожие; 3) Противоположный текст из подлинного текста, который был написан путем замены в оригинальном тексте некоторых слов на антонимы и фразы с противоположным значением; 4) Текст интерпретации из подлинного текста, который был написан как ответ на вопрос: Что вы думаете об определении метасистемы как системы?

Для алгоритмов сходства между цепочками текста были сравнены 4 текста по отношению к тексту-оригиналу, включая сравнение с самим текстом для оценки контроля, в результате получены результаты в виде процентного сходства между данными текстами.

Для систем определения плагиата “plagiarisma.net” и «Анти-плагиат», сначала был дан текст-оригинал с тем, чтобы убедиться, что указанные системы имеют оригинальный текст среди своих баз данных, затем были даны три оставшихся текста; эти системы дают процент оригинальности загруженного текста по отношению к совпадениям их сегментов с другими

существующими. Таким образом, что если процент оригинальности высок, сходство с текстом-оригиналом – низкий и наоборот.

Для метода, предложенного в данной работе, была проведена консультация с десятью экспертами в области информационных технологий, которым были даны слова и фразы из текста-оригинала вместе со списком из пяти возможных синонимов и не более двух антонимов или фраз с противоположным значением. Экспертам было предложено присвоить степень сходства указанных слов по шкале от 1 до 10. Слова принадлежат одному и тому же семантическому классу, который были выбраны из WordNet для русского языка. Для антонимов или фраз с противоположным значением было предложено выразить свое решение, признаются те, которые набрали более 60%. Промежуточные результаты, полученные для каждого слова, семантического класса считались степенью сходства.

Десять экспертов в области информационных технологий провели анализ четырех текстов, им было указано, что текст номер один - это текст-оригинал для сравнения с тремя остальными.

Варианты ответов были представлены в количественной шкале Лайкерта. Количественные результаты были преобразованы в качественные в процентной шкале, для сравнения их с результатами анализируемых методов, беря за образец результаты анализа экспертов. Полученные результаты и их сравнение с использованными методами и предложенным методом представлены и проанализированы ниже.

Результаты эксперимента в сравнении с другими методами приведены на рис.4. Что касается уровня сходства: 91% указали, что текст 2, по отношению к тексту 1, аналогичен или очень похож. 83% указали, что текст 3 значительно противоположен или абсолютно противоположен тексту 1. В то время как 75% утвердили, что текст 4 схож или схож в малой степени; что

переводится в проценты подобия таким образом: текст 2 = 84%; текст 3 = 82% и текст 4 = 42%.

В таком случае, как мы можем увидеть, предложенный метод для трех текстов имеет наиболее приближенное значение к мнениям экспертов, в том числе и для текста 2, в то время как другиеметоды дают отдаленные результаты или не определяют сходства.

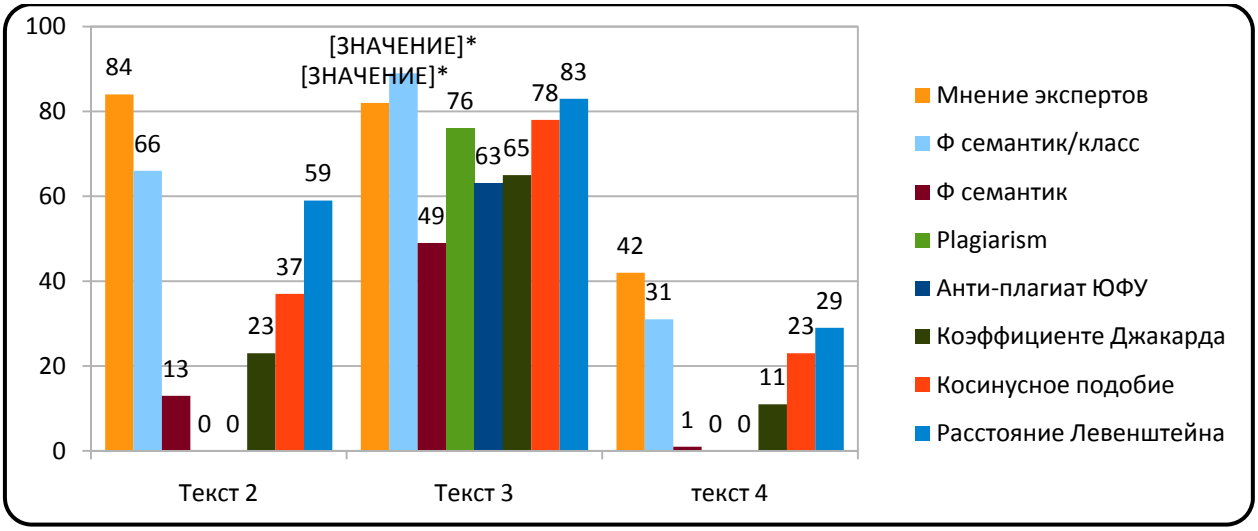

Рис. 4. – Результаты сходства текста.

Особого внимания заслуживают результаты, полученные для алгоритма расстояния Левенштейна, который имеетследующую особенность, еслитекстывводятсяскакими-тоизменениями порядка абзацев по отношению к фрагментам, результаты значительно уменьшаются. Например, при изменении абзаца три на абзац один текста 1, результаты сравнения снижаются на 37% для текста 2; 53% для текста 3 и 27% для текста. В то время как другие алгоритмы и методы сохраняют тот же самый процент. Это происходит в связи с тем, что алгоритм расстояния Левенштейна это минимальное количество операций, требуемых для трансформации одной цепочки характеристик на другую, и, при изменении порядка абзацев, увеличивается количество операций. Но изменение порядка абзацев одного текста не меняют его значения и тем более не могут замаскировать плагиат, в связи с этим, этот алгоритм неэффективен для целей сравнения.

Что касается определения плагиата по отношению к содержанию текста 1 - 75% указали, что текст 2 - это плагиат или плагиат с высокой долей процента. 67% указали, что текст 3 имеет высокий уровень плагиата, но с противоположным значением. В то время как 75% согласились с тем, что текст 4 - это не плагиат. Все результаты переводятся в следующие проценты плагиата: текст 2 = 73%, текст 3 = 89% и текст 4 = 13%. Указанные результаты сравниваются с результатами других методов на рис. 5. В нем можно проверить, что предложенный метод для всех трех текстов имеет результаты наиболее приближенные к мнению экспертов, в том числе и для текста 2, в то время как другие системы определения дают отдаленные результаты или не обнаруживают плагиата.

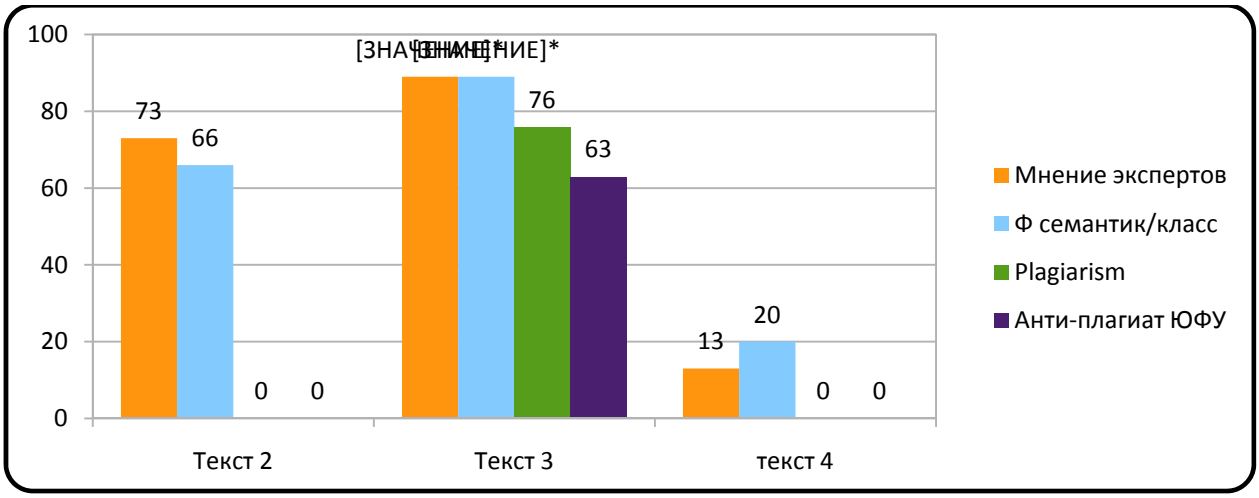

Рис. 5. - Результаты эксперимента «уровень плагиата»

Важно упомянуть, что в случае текста 3 эксперты указывают на противоположное значение по отношению к оригиналу, предлагаемый метод определяет сходство с отрицательным значением, в то время как сравниваемые системы обнаруживают только сходство. Поэтому на графиках значения мнения экспертов и предлагаемого метода обозначены символом *, и указанные значения не представлены, как отрицательноеотображение графика.

Выводы

Предлагаемый метод семантического сравнения между семантическими схемами текстовых пассажей позволяет сравнивать два текста, которые передают один смысл или противоположный смысл, которые написаны с использованием различной лексики, исключая совпадения в схожих текстовых пассажах, в отличие от существующих методов, которые лишь измеряют количество лексических компонентов, содержащихся в обоих текстах или максимальное значение сходства в парах слов.

Новые исследования в развитии этого предложения могут внести свой вклад в разработку методов для увеличения эффективности автоматической обработки текстов на естественном языке, в частности, в автоматическом сравнении сегментов семантически схожих текстов, написанных с использованием различной лексики, таких как в случае увеличения эффективности обнаружения модифицированного плагиата.

Литература

1. Maurer, H., Kappe, F., Zaka, B. Plagiarism - A Survey. Journal of Universal Computer Science. 2006. № 12 (8). pp. 1050-1084.

2. Bao, J-P., Shen, J-Y., Liu, X-D., Liu, H-Y., Zhang, X-D. Finding Plagiarism Based on Common Semantic Sequence Model // 5th International Conference on Advances, Daljan, China. 2004b. vol. 3129, pp. 640-645.

3. Bao, J-P., Shen, J-Y., Liu, X-D., Liu, H-Y., Zhang, X-D. Semantic Sequence Kin: A Method of Document Copy Detection // Advances in Knowledge Discovery and Data Mining, Sidney, Australia. 2004a vol. 3056, pp. 529-538.

4. Chi-Hong, L. и Yuen-Yan, C. A Natural Language Processing Approach to Automatic Plagiarism Detection // 8th ACM Conference on Information Technology Education, Flirida, USA. 2007. pp. 213–218.

5. Вишняков Р. Ю. Разработка и исследование формализованных представлений и семантических схем предложений текстов научно-

технического стиля для повышения эффективности информационного поиска: Дисс. канд. техн. наук: Таганрог. 2012. С. 92-106

6. Бермудес С. Х. Г. Подход к созданию модели семантического сравнения текстов. Информатизация и связь. Научно-практический журнал. 2016. №2. С. 121-126. Москва. ISSN: 2078-8320.

7. Бермудес С. Х. Г. О методе извлечения значимых текстовых пассажей как основы для сравнения текстов. Информатизация и связь. Научно-практический журнал. 2016. № 3. С. 213-219. Москва. ISSN: 2078-8320.

8. Salguero L. F. Resolución abductiva de anáforas pronominales. 2010. URL: personal.us.es/fsoler/papers/ivjornadas.pdf.

9. Agirre E., Cer D., Diab M., Gonzalez-Agirre A., WeiweiGuo. A pilot on semantic textual similarity // 6th International Workshop on Semantic Evaluation. 2012. pp. 385–393.

10. Agirre E., Cer D., Diab M., Gonzalez-Agirre A., Weiwei Guo. Semantic textual similarity // 2nd Joint Conference on Lexical and Computational Semantics, Атланта, USA. 2013. pp. 32–43.

11. Харламов А.А., Ермоленко Т.В., Дорохина Г.В. Сравнительный анализ организации систем синтаксических парсеров // Инженерный вестник Дона, 2013, №4.URL: ivdon.ru/ru/magazine/archive/n4y2013/2015.

12. Красников И.А., Никуличев Н.Н.Гибридный алгоритм классификации текстовых документов на основе анализа внутренней связности текста // Инженерный вестник Дона, 2013, №3 URL: ivdon.ru/ru/magazine/archive/n3y2013/1773.

13. Beltagy I. Chau C., Boleda G., Garrette D., Erk K., Mooney R. Montague meets markov: Deep semantics with probabilistic logical form // 2th Joint Conference on Lexical and Computational Semantics, Atlanta, USA. 2013. pp. 11-21.

14. Salehi B., Cook P. Predicting the compositionality of multiword expressions using translations in multiple languages// 2nd Joint Conference on Lexical and Computational Semantics, Atlanta, USA. 2013. pp. 266-275.

15. Palmer A., lexis Horbach A. и Pinkal M. Using the text to evaluate short answers for reading comprehension exercises // 2nd Joint Conference on Lexical and Computational Semantics, Atlanta, USA. 2013. pp. 520–524.

16. Resnik P. Using information content to evaluate semantic similarity in a taxonomy // 14th International Joint Conference on Artificial Intelligence, IJCAI'95, San Francisco, USA. 1995. pp. 448–453.

17. Mihalcea R., Corley C., Strapparava C. Corpus-based and knowledge-based measures of text semantic similarity // 21st National Conference on Artificial Intelligence, 2006. pp. 775–780.

18. Francesc Ll. C. Algoritmos de similitud entre cadenas de texto (php). 2015. URL: francescllorens.eu/00tokenizer/dst.php.

19. Рогозов, Ю. И. Подход к определению метасистемы как системы. Труды Института системного анализа РАН. 2013. № 4. С 92-110 URL: isa.ru/proceedings/images/documents/2013-63-4/t-4-13_92-110.pdf.

References

1. Maurer, H., Kappe, F. and Zaka, B. Plagiarism - A Survey. Journal of Universal Computer Science. 2006. № 12 (8). pp. 1050-1084.

2. Bao, J-P., Shen, J-Y., Liu, X-D., Liu, H-Y., Zhang, X-D. Semantic Sequence Kin: A Method of Document Copy Detection. Advances in Knowledge Discovery and Data Mining, Sydney, Australia. 2004a vol. 3056, pp. 529-538.

3. Bao, J-P., Shen, J-Y., Liu, X-D., Liu, H-Y., Zhang, X-D. Finding Plagiarism Based on Common Semantic Sequence Model. 5th International Conference on Advances, Daljan, China. 2004b. vol. 3129, pp. 640-645.

4. Chi-Hong, L., Yuen-Yan, C. A Natural Language Processing Approach to Automatic Plagiarism Detection. 8th ACM Conference on Information Technology Education (SIGITE'07), Florida, USA. 2007. pp. 213–218.

5. Vishnyakov R. YU. Razrabotka i issledovanief ormalizovannyh predstavlenij i semanticheskih skhem predlozhenij tekstov nauchno-tekhnicheskogo stilya dlya povysheniya ehffektivnosti informacionnogo poiska [Development and research of formal representations and semantic schemes of sentences of scientific and technical texts style to improve the efficiency of information retrieval]: Dyss. kand. tehn. nauk: Taganrog. 2012. pp. 92-106.

6. Bermudez S. J. G. Informatizaciya i svyaz. Nauchno-praktycheskyj zhurnal. 2016. Nº 2. pp.121-126. Moskva. ISSN: 2078-8320.

7. Bermudez S. J. G. Informatizaciya i svyaz. Nauchno-praktycheskyj zhurnal. 2016. Nº 3. pp. 213-219. Moskva. ISSN: 2078-8320.

8. Salguero L. F. Resolución abductiva de anáforas pronominales. 2010. URL: personal.us.es/fsoler/papers/ivjornadas.pdf.

9. Agirre E., Cer D., Diab M., Gonzalez-Agirre A., WeiweiGuo. A pilot on semantic textual similarity. 6th InternationalWorkshoponSemanticEvaluation (SemEval-2012). 2012. pp. 385–393.

10. Agirre E., Cer D., Diab M., Gonzalez-Agirre A., Weiwei Guo. Semantic textual similarity. 2nd Joint Conference on Lexical and Computational Semantics, Atlanta, USA. 2013. pp. 32–43.

11. Harlamov A.A., Ermolenko T.V., Dorohina G.V. Inženernyj vestnik Dona (Rus), 2013, № 4 URL: ivdon.ru/ru/magazine/archive/n4y2013/2015.

12. Krasnikov I.A., Nikulichev N.N. Inženernyj vestnik Dona (Rus), 2013, № 3 URL: ivdon.ru/ru/magazine/archive/n3y2013/1773.

13. Beltagy I. Chau C., Boleda G., Garrette D., Erk K., Mooney R. Montague meets markov: Deep semantics with probabilistic logical form. 2th Joint

Conference on Lexical and Computational Semantics, Atlanta, USA. 2013. pp. 11-21.

14. Salehi B., Cook P. Predicting the compositionality of multiword expressions using translations in multiple languages. 2nd Joint Conference on Lexical and Computational Semantics, Atlanta, USA. 2013. pp. 266-275.

15. Palmer A., lexis Horbach A., Pinkal M. Using the text to evaluate short answers for reading comprehension exercises. 2nd Joint Conference on Lexical and Computational Semantics, Atlanta, USA. 2013. pp. 520–524.

16. Resnik P. Using information content to evaluate semantic similarity in a taxonomy. 14th International Joint Conference on Artificial Intelligence, IJCAI'95, San - Francisco, USA. 1995. pp. 448–453.

17. Mihalcea R., Corley C., Strapparava C. Corpus-based and knowledge-based measures of text semantic similarity. 21st National Conferenceon Artificial Intelligence, 2006. pp. 775–780.

18. Francesc Ll. C. Algoritmos de similitud entre cadenas de texto (php). 2015- URL: francescllorens.eu/00tokenizer/dst.php.

19. Rogozov, YU. I. Trudi Ynstytuta systemnogo analyza RAN. 2013. № 4. pp. 92-110. URL: isa.ru/proceedings/images/documents/2013-63-4/t-4-13_92-110.pdf.