16. *Ermakov R.V., Kalihman D.M., L'vov A.A., Sokolov D.N.* Angular Velocity Estimation of Rotary Table Bench Using Aggregate Information from the Sensors of Different Physical Nature, *Proceedings of the 2017 IEEE Russia Section Young Researchers in Electrical and Electronic Engineering Conference (2017 ElConRus). February 1-3, 2017. St. Petersburg. Russia. 2017.*
17. *L'vov A.A., Mukhambetzhanov A.S.* Algoritm lokalizatsii spektral'nykh pikov [The algorithm of localization of spectral peaks], *Vestnik Saratovskogo gosudarstvennogo tekhnicheskogo universiteta* [Bulletin of Saratov state technical University], 2010, Vol. 4, No. 3 (51), pp. 154-156.
18. *Ermakov R.V., Kalikhman D.M., L'vov A.A.* Ispol'zovanie poligaussovskoy approksimatsii dlya opisaniya svoystv pogreshnostey opticheskogo datchika ugla [Use polygamously approximation to describe the properties of the errors in the optical angle sensor], *Trudy mezhdunarodnogo simpoziuma «Nadezhnost' i kachestvo»* [Proceedings of the international Symposium "Reliability and quality"], 2016, No. 2, pp. 23-25.
19. *Ermakov R.V., L'vov A.A.* Analiz pogreshnostey ugloizmeritel'nogo stenda na osnove opticheskogo beskontaktnogo datchika ugla [Error analysis of the angle measuring stand on the basis of contactless optical angle sensor], *Problemy upravleniya, obrabotki i peredachi informatsii: Cbornik trudov IV Mezhdunarodnoy nauchnoy konferentsii* [Control, processing and information transfer: proceedings of the IV International scientific conference]: in 2 vol. Vol. 2. Izd-vo: Rayt-Ekspo, 2015, pp. 116-123.
20. *Polushkin A.V., R.V. Ermakov R.V., Kaldymov N.A., et. al.* Algorithms, techniques and practical results of quality check automation of precision linear accelerometers, *21st Saint Petersburg International Conference on Integrated Navigation Systems, ICINS 2014 – Proceedings. 21,* 2014, pp. 213-219.

Статью рекомендовал к опубликованию д.т.н., профессор В.В. Курейчик.

Ермаков Роман Вячеславович – Саратовский государственный технический университет им. Гагарина Ю.А.; e-mail: roma-ermakov@yandex.ru; 410054, г. Саратов, ул. Политехническая, 77; тел.: +79053828180; аспирант.

Львов Алексей Арленович – e-mail: alvova@mail.ru; тел.: +79172015675; д.т.н.; профессор.

Светлов Михаил Семенович – Институт проблем точной механики и управления РАН (г. Саратов); e-mail: svetlovms@yandex.ru; 410028, г. Саратов, ул. Рабочая, 24; тел.: +79878263745; д.т.н.; в.н.с.

Ermakov Roman Vyacheslavovich – Yuri Gagarin State Technical University of Saratov; e-mail: roma-ermakov@yandex.ru; 77, Polytechnicheskaya street, Saratov, 410054, Russia; phone: +79053828180; postgraduate student.

L'vov Alexey Arlenovich – e-mail: alvova@mail.ru; phone: +79172015675; dr. of eng. sc.; professor.

Svetlov Michael Semenovich – Institute of Precision Mechanics and Control of RAS; e-mail: svetlovms@yandex.ru; 24, Rabochay street, Saratov, 410028, Russia; phone: +79878263745; dr. of eng. sc.; leading scientist.

УДК 004.89 DOI 10.23683/2311-3103-2017-3-17-29

Бермудес Сото Хосе Грегорио

МЕТОД ИЗМЕРЕНИЯ СЕМАНТИЧЕСКОГО СХОДСТВА ТЕКСТОВЫХ ДОКУМЕНТОВ

Рассматривается метод сравнения текстовых документов в обработке естественного языка на русском языке с целью определения их семантической близости; рассмотрим подзадачу измерения семантического сходства по критериям правильности и глубины. На основе проведённого обзора существующих подходов сравнения текстов, предложен

метод определения семантического подобия между двумя текстами на основе текстовых пассажей, который позволяет определить не только семантическую близость документов, представленных на естественном языке, но и дать количественную оценку сходства этих документов. Это исследование обрамлено в области автоматической обработки текстов (АОТ) и формализации естественных языков, постепенно переходя от самых простых методов анализа для более сложного, постепенно достигая уровень обработки, который уже можно увидеть текст не просто в виде последовательности слов, как единое целое, имеет некоторый смысл, так как оно соответствует человеческому восприятию. В соответствии с общей схемой автоматической обработки текста, данное исследование сосредоточено на семантическом уровне и представляет собой подробное описание заключительного этапа о сравнении на близость общей схемы. В основу метода положено определение степени подобия между текстовыми пассажами. Под текстовым пассажем будем понимать отдельное место в тексте, обладающее какой-то цельностью. Используется сегментация текстов, как основу для текстового сравнения в обработке естественного языка на русском языке; рассмотрим подзадачу извлечения фрагментов текста с особым смыслом, которые называются «текстовой пассаж». Также используется сравнение 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.

Измерение текстовой близости; определение подобия; сравнение текстов; представление семантических схем; текстовые пассажи.

Bermudez Soto José Gregorio

METHOD FOR MEASURING THE SEMANTIC-SIMILARITY OF TEXTUAL DOCUMENTS

The paper considers a method of comparing textual documents in the processing of natural language in Russian with the purpose of determining their semantic proximity; considered is the subtask of measuring the semantic similarity according to the criteria of correctness and depth. On the basis of the conducted review of existing approaches of texts comparison, a method is proposed for determining the semantic similarity between two texts on the basis of textual passages, which makes it possible to determine not only the semantic proximity of documents presented in natural language, but also quantify the similarity of these documents. This study is framed in the field of automatic text processing (ATP) and the formalization of natural languages, gradually shifting from the simplest methods of analysis to the more complex, gradually reaching a level of processing that can already see the text not just as a sequence of words, but as a single whole, which has some meaning, as it corresponds to human perception. In accordance with the general scheme of automatic text processing, this study is focused on the semantic level and is a detailed description of the final stage about comparing the closeness of the general scheme. The method is based on determining the degree of similarity between the passages. Under the passage we mean a separate place in the text, which has some kind of integrity. This work uses segmentation of texts as a basis for text comparison in the natural language processing in Russian; it will be considered subtask of extracting parts of text with a special meaning, which are called "passage". Also the comparison of texts in Russian is used, in the subtask of determination of semantic proximity. A review of existing methods of comparison is given. The determination method of degree of similarity between textual passages within a semantic class is proposed. Existing methods are compared with the proposed method and a comparison made by people in an experiment, which shows the suitability of the proposed method.

Measurement of textual proximity; definition of similarity; comparison of texts; presentation of semantic schemes; passages.

Введение. Все интеллектуальные системы обработки текста на естественном языке, на современном этапе, нашли свое применение в различных сферах и решают конкретные задачи. К системам, используемым АОТ можно отнести: издательские системы [1]; автоматизированные лексикографические системы подготовки и использования словарей [2]; информационно-поисковые системы [3]; системы машинного перевода (МП) [4, 5]; системы понимания и распознавания речи [6] и другие.

Большинство исследований автоматической обработки естественного языка направлены на приложения и системы, которые могут быть сгруппированы на системы Автоматической Категоризации Текстов [7], системы антиплагиата и системы реферирования среди прочих; в которых текстовое сравнение, в зависимости от обстоятельств, рассматривается в качестве второстепенной задачи. Кроме того, когда проводятся исследования именно по теме текстового сравнения, оно проводится для целей систем информационного поиска и антиплагиата. В связи с этим не проводится глубоких исследований семантического содержания обрабатываемых текстов.

В настоящее время поиск сходства между текстами имеет большое практическое применение, в том числе для обнаружения плагиата в научных и творческих исследованиях. В работе [8] 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 [9–11]; так и для поиска информации, так как это отображено в работе [12].

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 [13], выходом этого первого уровня будет перечень значимых пассажей из каждого документа, которые послужат в качестве входных данных для следующего уровня для разрешения анафоры [14] а последний, в свою очередь, поступает на уровень семантического представления схем [12]. Построенная схема представления является входом для обнаружения уровня семантического сходства между текстовыми пассажами.

На этом уровне, в качестве входных данных используются текстовые пассажи из документов и на основе их результатов сравнения, выполняется вычисление сходства между этими двумя документами.

1. Обзор и постановка задачи. Нахождение семантического сходства между парами текстов является серьёзной проблемой для АОТ. Такая проблема возникает в различных задачах АОТ, таких как машинный перевод, автоматическое построение рефератов, определение авторства, обнаружение академического плагиата, тест на понимание текста, поиска информации; и многие другие, в которых нужно определить и измерить степень подобия между двумя заданными текстами.

Поиск степени семантической схожести текстов был рассмотрен в качестве одной из основных задач в рамках многих российских и международных конференций [15, 16], чему уделяется значительное внимание в последние годы. Многие из разработанных и применяемых 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 исследовании [17], была представлена попытка создания семантической модели для выявления неявных аргументов в текстах. Здесь утверждается, что, несмотря на то, что это является лёгкой задачей для читателя-человека, это является трудной для компьютеров, потому что нет никакого способа сообщить им, что аргумент может быть выведен несколько раз в тексте.

Цель текстуального семантического сходства, захватить когда смысл двух текстов является сходным. Эта концепция шире, чем найти степень текстуального сходства, как и в случае вышеназванных алгоритмов, которые измеряют только количество терминов обоих текстов, но не измеряет степень сходства этих текстов. Причём, сходство должно быть выражено конкретным значением.

В целом, концепция релевантности информации основана на её количественной оценке. Схематически это можно представить следующим образом: имеется документ D и запрос Q, конечной целью является измерить сходство или релевантность между обоими.

$$sim(D, Q)=? \quad (1)$$

Для того, чтобы определить упомянутую релевантность, системы поиска информации непосредственно применяют ряд функций, которые называются меры сходства, которые количественно оценивают релевантность между документом и запросом.

По сути эти меры основываются на количестве терминов, которые совместно встречаются как в документе, так и в запросе.

Таким образом, для предложенного метода, в отличии от поисковых систем, конечной целью будет являться определение сходства или релевантности между двумя текстами D_1 и D_2:

$$sim(D_1, D_2)=? \quad (2)$$

То есть, вычислить сходство между документами D_1 и D_2 в функции сходства между соответствующими текстовыми пассажами

$$sim(D_1,D_2)= f(X_{ij} .. X_{nm}). \quad (3)$$

Как было отмечено выше, существуют также методы, которые вычисляют сходство между двумя документами посредством алгоритмов, основанных на частоте встречаемости терминов внутри обоих документов, но это не более чем методы, которые распространяется на сравнение запроса и документа, в котором один из них является запросом. Таким образом, достаточно проверить вычисление сходства между запросом и документом.

Мера сходства позволяет определить сходство между двумя сегментами текста (будь это целый документ или пассаж сам по себе) и запрос, или в нашем случае между двумя текстовыми пассажами. Традиционно эти меры основываются, главным образом, на терминах, существующих в обоих текстах и в запросе, и также на дискриминационном значении каждого термина.

Методы информационного поиска реализуют эти вычисления сходства, определяя документ D как набор пар значений (d_i, n_i), в которых d_i – это термин, n_i – это количество повторений указанного термина в документе. Значение N представляет размер документа в соответствии количеству терминов, которые его формируют, таким образом:

$$D = ((d_1, n_1), (d_2, n_2), \ldots (d_N, n_N)). \quad (4)$$

С другой стороны, в этом же самом подходе представления, значение Q определяется как набор пар значений (q_i, m_i), в которых q_i – это термин и m_i – количество указанных терминов в вопросе. Значение K указывает количество терминов отличных от запроса, таким образом:

$$Q = ((q_1, m_1), (q_2, m_2), \ldots (q_K, m_K)). \quad (5)$$

Мера значения сходства между Q и D вычисляется, среди прочих методов, в соответствии с:

- Количеством терминов, которые встречаются как в запросе, так и в документе.
- Количеством появлений в обоих (запросе и документе) указанных общих появлений терминов.
- Дискриминантное значение или вес x_i термина внутри собрания документов. Этот вес x_i одного термина t_1, определяется в соответствии с количеством документов собрания, в которых появляется указанный термин.

Таким образом, мера сходства вычисляется в соответствии:

$$sim(D, Q) = Y\forall_{i\in Q\wedge D}\ (t_i, n_i, m_i, x_i, N). \quad (6)$$

где Y – метод для определения значения сходства между документом и запросом в соответствии с параметрами.

В других методах, которые используют текстовые пассажи, в качестве единицы обработки, вычисление сходства между текстовыми пассажами такое же подход, но появления терминов заменяются пассажами для того, чтобы затем вычислить сходство между запросом и документом в соответствии со сходством всех текстовых пассажей. К тому же, во многих из них нет чёткой методики, сегментации документа на текстовые пассажи и вычисления меры сходства между документами.

Методы сравнения текстов. Исследование [18], 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 создание статистического словаря. Если перевод возможен, это означает, что термин эквивалентен термину в тексте, выраженному другими словами.

Другой способ подойти к этой задаче – это рассмотреть её как проблему Question Answering, где один из текстов является вопросом, а другой ответом. Это суть работы [19], где предлагается модель, которая измеряет степень сходства в функции, если ответ действительно отвечает на вопрос.

Особо следует упомянуть процедуру, указанную в работе [12], где возникает сравнение семантического сходства фрагментов текста и строится семантический критерий сравнения, учитывающий структуру семантических схем, в этом смысле автор объясняет, что: "*Пусть s_a и s_t семантические схемы фрагментов текстов q и t соответственно. Тогда критерий близости φ данных семантических схем определим следующим образом*":

$$\varphi(S_q,S_t) = (S_q \approx S_t);$$
$$\varphi(S_q,S_t) \in D;$$
$$D=[0..1].$$

где символ ≈ обозначает операцию установления близости, а D – 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 Вишнякова Р.Ю.

Базовые критерии сравнения близости. В которых 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 методы. В которых 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 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; расстояние Левенштейна, которое состоит из минимального количества необходимых операций для трансформации одной цепочки характеристик в другую.

Текстуальное семантическое сходство имеет своей целью уловить момент, когда смысл двух текстов аналогичен. Это понятие шире, чем найти степень текстуального подобия, как и в случае выше упомянутых алгоритмов, они измеряют только количество лексических компонентов, которые разделяют оба текста, то есть, которые не измеряет сходство двух текстов относительно значения, которое должно быть выражено.

Вторая группа характеристик рассмотрена так же в исследованиях: Leacock & Chodorow [20], Lesk [21], Wu & Palmer [22], Resnik [23] , Lin [24], и Jiang & Conrath [25], 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 [26]. Использование взаимной информации (PMI) [27] для вычисления подобия между парами слов, и латентно-семантического анализа (ЛСА) [28].

2. Предлагаемый метод. Следует отметить, что в основе предлагаемого метода, фрагменты текстов уже являются текстовыми пассажами с семантическим содержанием [13], а не любые фрагменты.

На уровне представления текстовых пассажей в семантических схемах получается число n-схем пассажей эталона и число m-схем пассажей сравниваемого текста, которые в последствии будут сравнены в соотношении n: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.

Большинство методов информационного поиска вычисляют сходство документа в соответствии со сходством их текстовых пассажей, в которых функция Y, выражение (6), может быть по существу наиболее схожий текстовой пассаж или сумма сходств.

В нашем случае ситуация иная, в связи с тем, что целью сравнения являются два документа. В связи с этим, точность будет зависеть от правильности и полноты сравниваемого текста по отношению к эталону, который зависит напрямую от цели сравнения, и как следствие, оценки. Таким образом, в нашем случае, чтобы определить сходство между документами, будут применяться оба подхода.

1. Предыдущее связано, в основном, с тем, что сравнивая один текст с другим, каждый текстовой пассаж сравниваемого текста должен быть сопоставлен со всеми пассажами текста-эталона в соотношении n:m; из которого будут выбраны текстовые пассажи с большим сходством. Соответственно:

$$sim(P_1, P_2) = max \forall ij \in P_1 \wedge P_2\ sim(P_{1i}, P_{2j}). \quad (7)$$

Таким образом, что указанное сходство отделяет как текстовой пассаж сравниваемого текста, так и текстовой пассаж текста-эталона.

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.

$$\Phi_{\text{семантик/класс}} = \frac{\sum_i^k \frac{\sum p_j}{l}}{n}, \quad (8)$$

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.

Определение правильности и глубины напрямую зависит от целей и задач сравнения, а, следовательно, и оценки. Жизнеспособный критерий правильности вытекает из результатов, полученных на предыдущем этапе, но теперь по отношению ко всему тексту, то есть коэффициент сходства между эталоном и сравниваемым текстом – C, определяется по формуле:

$$C = \frac{\sum_1^q \Phi_i}{m}, \quad (9)$$

где Φ_i – результат, полученный для каждого i-того сравнения; q – количество текстовых пассажей сравниваемого текста; и m – общее число текстовых пассажей эталона.

Глубина сравнения может быть определена в пропорциональности количества текстовых пассажей сравниваемого текста, по отношению к количеству текстовых пассажей эталона, то есть коэффициент глубины S, определяется по формуле:

$$S = \frac{q}{m}, \quad (10)$$

В то время как оценка может быть обозначена средним арифметическим двух ранее полученных значений C и S, то итоговая оценка сходства между документами R, может быть определена по формуле:

$$R = \frac{C+S}{2}, \quad (11)$$

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 [29].

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 «Антиплагиат», 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 [12], который называется «*Ф семантик*».

Для проведения эксперимента были выбраны четыре текста: 1) введение научной статьи под названием «Подход к определению метасистемы как системы»; 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. 1. Что касается уровня сходства: 91 % указали, что текст 2, по отношению к тексту 1, аналогичен или очень похож. 83 % указали, что текст 3 значительно противополо-

жен 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.

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. 2. 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.

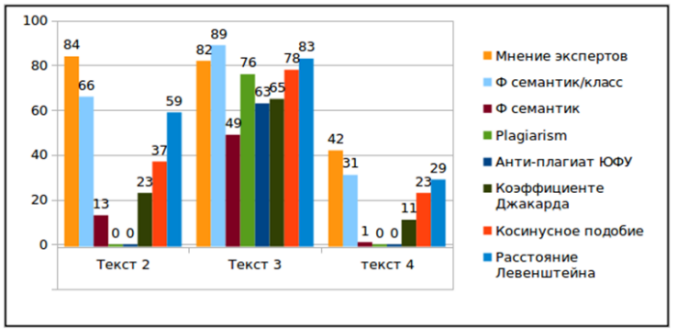

Рис. 1. Результаты сходства текста

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.

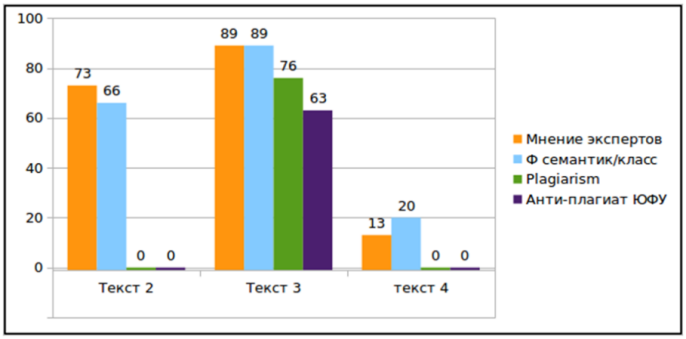

Рис. 2. Результаты эксперимента «уровень плагиата»

Выводы. Таким образом, в работе предложен и обоснован метод сравнения текстов на уровне представления текстовых пассажей в семантических схемах, как наиболее эффективный по сравнению с имеющимися в настоящее время. Кроме того, данный метод позволяет определить не только семантическую близость документов, представленных на естественном языке, но и дать количественную оценку сходства этих документов, что является положительным результатом и может широко использоваться в практических задачах. 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.

БИБЛИОГРАФИЧЕСКИЕ ССЫЛКИ

1. Языкознание. Бол. энцикл. словарь / гл. ред. В.Н. Ярцева. – 2-е изд. – М.: Бол. рос. энцикл., 1998. – 685 с.
2. *Марчук Ю.Н.* Компьютерная лингвистика. – М.: АСТ; Восток-Запад, 2007. – 317 с.
3. *Гайдамакин Н.А.* Автоматизированные информационные системы, базы и банки данных. Вводный курс: учеб. пособие. – М.: Гелиос АРВ, 2002. – 368 с.
4. *Баранов А.Н.* Введение в прикладную лингвистику: учеб. пособие. – М.: Эдиториал УРСС, 2001. – 360 с.
5. Искусственный интеллект. В 3 кн. Кн. 1. Системы общения и экспертные системы: справочник / под ред. Э.В. Попова. – М.: Радио и связь, 1990. – 464 с.
6. *Потапова Р.К.* Речь: коммуникация, информация, кибернетика: учеб. пособие. – М.: Едиториал УРСС, 2003. – 568 с.
7. *Muñoz T.R.* Representación del conocimiento textual mediante técnicas lógico-conceptuales en aplicaciones de tecnologías del lenguaje humano // Tesis doctoral. Universidad de Alicante. – España, 2009. – 128 p.
8. *Maurer H., Kappe F. y Zaka B.* Plagiarism – A Survey // Journal of Universal Computer Science. – 2006. – No. 12. – P. 1050-1084.
9. *Bao J-P., Shen J-Y., Liu X-D., Liu H-Y. y Zhang X-D.* Semantic Sequence Kin: A Method of Document Copy Detection // Advances In Knowledge Discovery and Data Mining. Lecture Notes in Artificial Intelligence (LNAI) – Sydney, Australia, 2004. – Vol. 3056. – P. 529-538.
10. *Bao J-P., Shen J-Y., Liu X-D., Liu H-Y. y Zhang X-D.* Finding Plagiarism Based on Common Semantic Sequence Model // The 5th International Conference on Advances in Web-Age Information Management (WAIM). Lecture Notes in Computer Science – China: Dalian, 2004. – Vol. 3129. – P. 640-645.
11. *Chi-Hong L. y Yuen-Yan C.* A Natural Language Processing Approach to Automatic Plagiarism Detection // The 8th ACM Conference on Information Technology Education (SIGITE'07) – Florida, USA, 2007. – P. 213-218.
12. *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.
13. *Бермудес С.Х.Г.* О методе извлечения значимых текстовых пассажей как базы для текстового сравнения // Информатизация и и связь. – 2016. – № 3. – С. 231-219.
14. *Salguero L.F.* Resolución abductiva de anáforas pronominales. – http://www.http://personal.us.es/fsoler/papers/ivjornadas.pdf. (дата обращения 29.01.2016).
15. *Agirre E., Cer D., Diab M., Gonzalez-Agirre A. and Weiwei Guo.* A pilot on semantic textual similarity // The 6th International Workshop on Semantic Evaluation (SemEval-2012 task 6) – Atlanta, USA, 2012. – P. 385-393.
16. *Agirre E., Cer D., Diab M., Gonzalez-Agirre A. and Weiwei Guo.* Semantic textual similarity // 2nd Joint Conference on Lexical and Computational Semantics (*SEM-2013) – Georgia, USA, 2013. – P. 32-43.

17. *Michael R. and Anette F.* Automatically identifying implicit arguments to improve argument linking and coherence modeling // 2nd Joint Conference on Lexical and Computational Semantics (*SEM-2013) – Georgia, USA, 2013. – P. 321-333.
18. *Salehi B. and Cook P.* Predicting the compositionality of multiword expressions using translations in multiple languages // Second Joint Conference on Lexical and Computational Semantics (*Sem-2013), Atlanta, Georgia, USA. 2013. – P. 134-142.
19. *Palmer A., lexis Horbach A. and Pinkal M.* Using the text to evaluate short answers for reading comprehension exercises // Second Joint Conference on Lexical and Computational Semantics (*SEM). – Vol. 2 (SemEval 2013) – Atlanta, Georgia, USA, 2013. – P. 520-524.
20. *Leacock C. and Chodorow M.* Combining local context and wordnet similarity for word sense identification // Christiane Fellbaum, editor, MIT Press, 1998. – P. 265-283.
21. *Lesk M.* Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone, *5th Annual International Conference on Systems Documentation*, 1986, pp. 24-26. ACM.
22. *Wu Zhibiao and Stone Palmer M.* Verb semantics and lexical selection // James Pustejovsky, editor, ACL, 1994. – P. 133-138. Morgan Kaufmann Publishers / ACL
23. *Resnik P.* Using information content to evaluate semantic similarity in a taxonomy // 14th International Joint Conference on Artificial Intelligence, IJCAI'95. – San Francisco, CA, USA, 1995. – P. 448-453.
24. *Lin Dekang.* An information-theoretic definition of similarity // Fifteenth International Conference on Machine Learning, ICML '98. – San Francisco, CA, USA. Morgan Kaufmann Publishers Inc., 1998. – P. 296-304.
25. *Jiang Jay J. and Conrath D.W.* Semantic similarity based on corpus statistics and lexical taxonomy // 10th International Conference on Research in Computational Linguistics, ROCLING'97. – 1997. – P. 19-33.
26. *Mihalcea R., Corley C. and Strapparava C.* Corpus-based and knowledge-based measures of text semantic similarity // 21st National Conference on Artificial Intelligence. – 2006. – P. 775-780.
27. *Turney Peter D.* Mining the web for synonyms: Pmi-ir versus lsa on toefl // 12th European Conference on Machine Learning. – 2001. – P. 491-502.
28. . *Landauer Thomas K., Foltz Peter W. and Laham Darrell.* An Introduction to Latent Semantic Analysis. Discourse Processes. – Springer-Verlag, 1998. – P. 259-284.
29. *Francesc Ll. C.* Algoritmos de similitud entre cadenas de texto (php). – 2015. – URL: francescllorens.eu/00tokenizer/dst.php.

REFERENCES

1. Yazykoznanie. Bol. entsikl. slovar' [Linguistics. Big encyclopaedic dictionary], chief ed. V.N. Yartseva. 2nd ed. Moscow: Bol. ros. entsikl., 1998, 685 p.
2. *Marchuk Yu.N.* Komp'yuternaya lingvistika [Computational linguistics]. Moscow: AST; Vostok-Zapad, 2007, 317 p.
3. *Gaydamakin N.A.* Avtomatizirovannye informatsionnye sistemy, bazy i banki dannykh. Vvodnyy kurs: ucheb. posobie [Automated information systems, databases and data. Introductory course: textbook]. Moscow: Gelios ARV, 2002, 368 p.
4. *Baranov A.N.* Vvedenie v prikladnuyu lingvistiku: ucheb. posobie [Introduction to applied linguistics: textbook]. Moscow: Editorial URSS, 2001, 360 p.
5. Iskusstvennyy intellect [Artificial intelligence]. In 3 book. Book 1. *Sistemy obshcheniya i ekspertnye sistemy: spravochnik* [Communication and expert systems: a Handbook], ed. byE.V. Popova. Moscow: Radio i svyaz', 1990, 464 p.
6. *Potapova R.K.* Rech': kommunikatsiya, informatsiya, kibernetika: ucheb. posobie [Speech: communication, information, cybernetics: textbook]. Moscow: Editorial URSS, 2003, 568 p.
7. *Muñoz T.R.* Representación del conocimiento textual mediante técnicas lógico-conceptuales en aplicaciones de tecnologías del lenguaje humano, *Tesis doctoral. Universidad de Alicante*. España, 2009, 128 p.
8. *Maurer H., Kappe F. y Zaka B.* Plagiarism – A Survey, *Journal of Universal Computer Science*, 2006, No. 12, pp. 1050-1084.

9. *Bao J-P., Shen J-Y., Liu X-D., Liu H-Y. y Zhang X-D.* Semantic Sequence Kin: A Method of Document Copy Detection, *Advances In Knowledge Discovery and Data Mining. Lecture Notes in Artificial Intelligence (LNAI) – Sydney, Australia, 2004,* Vol. 3056, pp. 529-538.
10. *Bao J-P., Shen J-Y., Liu X-D., Liu H-Y. y Zhang X-D.* Finding Plagiarism Based on Common Semantic Sequence Model, *The 5th International Conference on Advances in Web-Age Information Management (WAIM). Lecture Notes in Computer Science – China: Dalian, 2004,* Vol. 3129, pp. 640-645.
11. *Chi-Hong L. y Yuen-Yan C.* A Natural Language Processing Approach to Automatic Plagiarism Detection, *The 8th ACM Conference on Information Technology Education (SIGITE'07) – Florida, USA, 2007,* pp. 213-218.
12. *Vishnyakov R.Yu.* Razrabotka i issledovanie formalizovannykh predstavleniy i semanticheskikh skhem predlozheniy tekstov nauchno-tekhnicheskogo stilya dlya povysheniya effektivnosti informatsionnogo poiska: diss. … kand. tekhn. nauk [The development and study of formal representations and semantic diagrams of the sentences of the texts of scientific-technical style to improve the efficiency of information retrieval. Cand. of eng. sc. diss.]. Taganrog, 2012.
13. *Bermudes S.Kh.G.* O metode izvlecheniya znachimykh tekstovykh passazhey kak bazy dlya tekstovogo sravneniya [On the method of extraction of important text passages as a basis for tech-stowage comparison], *Informatizatsiya i i svyaz'* [Informatization and communication], 2016, No. 3, pp. 231-219.
14. *Salguero L.F.* Resolución abductiva de anáforas pronominales. Available at: http://www.http://personal.us.es/fsoler/papers/ivjornadas.pdf. (accessed 29 January 2016).
15. *Agirre E., Cer D., Diab M., Gonzalez-Agirre A. and Weiwei Guo.* A pilot on semantic textual similarity, *The 6th International Workshop on Semantic Evaluation (SemEval-2012 task 6) – Atlanta, USA, 2012,* pp. 385-393.
16. *Agirre E., Cer D., Diab M., Gonzalez-Agirre A. and Weiwei Guo.* Semantic textual similarity, *2nd Joint Conference on Lexical and Computational Semantics (*SEM-2013) – Georgia, USA, 2013,* pp. 32-43.
17. *Michael R. and Anette F.* Automatically identifying implicit arguments to improve argument linking and coherence modeling, *2nd Joint Conference on Lexical and Computational Semantics (*SEM-2013) – Georgia, USA, 2013,* pp. 321-333.
18. *Salehi B. and Cook P.* Predicting the compositionality of multiword expressions using translations in multiple languages ,*Second Joint Conference on Lexical and Computational Semantics (*Sem-2013), Atlanta, Georgia, USA. 2013,* pp. 134-142.
19. *Palmer A., lexis Horbach A. and Pinkal M.* Using the text to evaluate short answers for reading comprehension exercises, *Second Joint Conference on Lexical and Computational Semantics (*SEM). – Vol. 2 (SemEval 2013) – Atlanta, Georgia, USA, 2013,* pp. 520-524.
20. *Leacock C. and Chodorow M.* Combining local context and wordnet similarity for word sense identification, *Christiane Fellbaum, editor, MIT Press,* 1998, pp. 265-283.
21. *Lesk M.* Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone, *5th Annual International Conference on Systems Documentation,* 1986, pp. 24-26. ACM.
22. *Wu Zhibiao and Stone Palmer M.* Verb semantics and lexical selection, James *Pustejovsky, editor, ACL, 1994,* pp. 133-138. Morgan Kaufmann Publishers / ACL.
23. *Resnik P.* Using information content to evaluate semantic similarity in a taxonomy, *14th International Joint Conference on Artificial Intelligence, IJCAI'95.* San Francisco, CA, USA, 1995, pp. 448-453.
24. *Lin Dekang.* An information-theoretic definition of similarity, *Fifteenth International Conference on Machine Learning, ICML '98. – San Francisco, CA, USA. Morgan Kaufmann Publishers Inc., 1998,* pp. 296-304.
25. *Jiang Jay J. and Conrath D.W.* Semantic similarity based on corpus statistics and lexical taxonomy, *10th International Conference on Research in Computational Linguistics, ROCLING'97,* 1997, pp. 19-33.

26. *Mihalcea R., Corley C. and Strapparava C.* Corpus-based and knowledge-based measures of text semantic similarity, *21st National Conference on Artificial Intelligence*, 2006, pp. 775-780.
27. *Turney Peter D.* Mining the web for synonyms: Pmi-ir versus lsa on toefl, *12th European Conference on Machine Learning*, 2001, pp. 491-502.
28. . *Landauer Thomas K., Foltz Peter W. and Laham Darrell.* An Introduction to Latent Semantic Analysis. Discourse Processes. Springer-Verlag, 1998, pp. 259-284.
29. *Francesc Ll. C.* Algoritmos de similitud entre cadenas de texto (php). 2015. Available at: francescllorens.eu/00tokenizer/dst.php.

Статью рекомендовал к опубликованию д.т.н., профессор В.И. Финаев.

Бермудес Сото Хосе Грегорио – Южный федеральный университет; e-mail: jbermudesoto@gmail.com; 347928, г. Таганрог, ул. Энгельса, 1; тел.: +79885792533; аспирант.

Bermudez Soto José Gregorio – Southern Federal University; e-mail: jbermudesoto@gmail.com; 1, Engel'sa street, Taganrog, 347928, Russia; phone: +79885792533; postgraduate student.

УДК 621.315:621.317.7:621.391 DOI 10.23683/2311-3103-2017-3-29-42

С.А. Кузин, П.А. Львов, А.А. Львов, М.С. Светлов

ПОВЫШЕНИЕ ТОЧНОСТИ ЕМКОСТНЫХ ДАТЧИКОВ ДАВЛЕНИЯ ДЛЯ АВИАКОСМИЧЕСКОЙ ТЕХНИКИ

Работа посвящена разработке новых емкостных датчиков абсолютного и избыточного давления, используемых для нужд авиационной и космической техники и удовлетворяющих требованиям программы импортозамещения. Предложена методика повышения точности измерения современных цифровых интеллектуальных емкостных датчиков давления, которая основана на использовании нового формирователя сигналов датчиков, когда чувствительный элемент и опорная емкости включаются в петлю переменного тока, и обработке выходных оцифрованных сигналов по методу максимального правдоподобия. В отличие от известного способа построения формирователя резистивных датчиков на основе петли постоянного тока предложено использовать генератор переменного тока в качестве источника опорного сигнала. Рассмотрена математическая модель системы: чувствительный элемент–формирователь сигнала, показано, как за счет ее усложнения можно снизить требования к используемым источнику переменного тока и операционным усилителям. Задача оценивания неизвестных параметров полученной математической модели сведена к решению линейной системы уравнений с билинейным ограничением на вновь введенные неизвестные параметры, приведены выражения для оценок параметров математической модели датчика, получаемых в результате применения итеративной процедуры. Эти оценки обладают всеми оптимальными свойствами оценок максимального правдоподобия. Обсуждаются достоинства предлагаемой методики, среди которых можно выделить высокие точность измерения и чувствительность датчика, а также более простую конструкцию формирователя сигналов и относительно невысокую себестоимость датчика. Проведен сравнительный анализ достигаемой точности измерения предлагаемой методики и двух классических известных методов измерения с помощью имитационного моделирования. Показано, что точность нового датчика примерно на порядок превосходит точность существующих аналогов, производимых в настоящее время.

Интеллектуальный цифровой датчик; емкостной датчик давления; петля переменного тока; формирователь сигналов; оптимальное оценивание; метод максимального правдоподобия.