Метод выявления заимствований в текстах разноязычных документов

© В. Н. Захаров © Ал-др А. Хорошилов © Ал-ей А. Хорошилов

ФИЦ ИУ РАН,
Москва
VZakharov@ipiran.ru Khoroshilov@mail.ru A.A.Horoshilov@mail.ru

Аннотация

В работе рассматривается метод автоматического выявления заимствований в текстах разноязычных документов, основанный на сопоставлении их формализованных представлений. При решении данной задачи была разработана модель представления смысловой структуры текстов и методы формализации и установления смысловой близости между фрагментами сравниваемых разноязычных текстов. Основным преимуществом данного метода является то, что он позволяют эффективно выявить различного рода заимствования, включая более сложные случаи плагиата.

Статья подготовлена при частичной поддержке гранта РФФИ 16-07-01028.

1 Введение

1.1 Проблема выявления заимствований в текстах документов

Наличие заимствований в работах, относящихся к сфере образования и науки, является на данный момент серьезной проблемой во многих странах мира. В связи с этим в зарубежной академической практике западных университетов и научных журналов существуют документы, регулирующие правила заимствований текста и оформления соответствующих ссылок на источники, а также четко прописаны критерии отнесения некорректных заимствований к плагиату в различных формах. Плагиатом, как правило, считается любое использование чужих идей и высказываний без должной отсылки к источнику. Заимствованием также считается пересказ текста другого источника, не сопровождающийся указанием на источник заимствования идей. В нашей стране, к сожалению, критерии выявления плагиата регламентированы не столь серьезно. Но во многих ведущих ВУЗах введены положения, которые подробно определяют ответственность учащихся за любые виды заимствований в своих работах. Для выявления заимствований во многих учреждениях образования и науки функционируют специальные информационные системы. К сожалению, возможности этих систем серьезно ограничены и они не позволяют выявлять заимствования при существенном изменении недобросовестным автором лексического состава или структуры исходного текста, а также заимствования из текстов, представленных на другом языке.

1.2 Обзор существующих подходов к задаче выявления заимствований в текстах разноязычных документов

В настоящее время задача выявления заимствований в текстах разноязычных документов недостаточно изучена в нашей стране. Поэтому не существует инструментария, позволяющего выявлять заимствования из иностранной литературы. В то же время в работах иностранных ученых эта проблема активно изучается. Так в работе [1] авторы сводят процесс поиска плагиата к трем этапам: 1) Поиск документов-кандидатов. Для этого документ автоматически переводится. Затем из документа извлекаются ключевые слова, которые после этого используются для поиска документов-кандидатов. 2) Подробный анализ документов-кандидатов. Для этого могут использоваться три поисковые модели: модель 3-грамм; явная модель семантического анализа, модель анализа подобия на основе межъязыкового выравнивания. На основе использования данных моделей принимается решение о наличии в документах-кандидатах плагиата. 3) Документы-кандидаты подробно анализируются для того, чтобы выявлять случаи, когда найденные заимствования не являются плагиатом, например, если скопированные разделы являются цитатами.

В работе [2] авторы предлагают разделить процесс поиска плагиата на 4 этапа: 1) фаза предварительной обработки (разбиение на лексемы, удаление стоп-слов); 2) извлечение ключевых слов и перевод; 3) выбор документов-кандидатов; 4) поиск

Труды XVIII Международной конференции DAMDID/RCDL'2016 «Аналитика и управление данными в областях с интенсивным использованием данных», Ершово, 11-14 октября 2016

плагиата с помощью методов, используемых для одноязычных текстов. Данный метод был разработан для сопоставления текстов на арабском и английском языках. Эксперимент показал довольно высокие показатели полноты и точности.

В работе [3] авторы предлагают метод под названием MLPlag, основанный на анализе местоположения слов. В данной работе используется тезаурус EuroWordNet для формирования независимого от языка представления текста. Детальное сравнение текстов проводится путем вычисления симметричных и асимметричных мер подобия.

Рассмотренные и другие схожие методы [1-11], разработанные зарубежными учеными, демонстрируют основные тенденции решения задачи выявления заимствований в текстах разноязычных документов. Основным недостатком, который присутствует во всех этих работах, на наш взгляд, является попытка разделять документ на отдельные слова, которые затем авторы методов пытаются перевести отдельно от контекста. Такой подход может привести к значительному числу ошибок.

2 Выявление заимствований в текстах разноязычных документов

2.1 Теоретическое представление о смысловой структуре текста

В качестве базовой теоретической концепции при разработке метода выявления заимствований в текстах разноязычных документов использовалась концепция проф. Г.Г. Белоногова и проф. Р.С. Гиляревского, констатирующая, что смысловое содержание текстов выражается с помощью единиц смысла, входящих в их состав. По их мнению, наиболее устойчивыми единицами смысла являются понятия. Проф. Г.Г. Белоногов определяет термин «понятие» как «социально значимый мыслительный образ, за которым в языке закреплено его наименование в виде отдельного слова или, значительно чаще, в виде устойчивого фразеологического словосочетания…» [14,18,27]. Понятия занимают центральное место в языке и речи и являются теми базовыми строительными блоками, на основе которых формируются смысловые единицы более высоких уровней.

Также при разработке метода были использованы конструктивные признаки текста: глобальная и локальная связности текстов [16,17,18]. Глобальная связность обеспечивает раскрытие темы документа, а локальная связность проявляется во взаимосвязи между соседними единицами текста. В соответствии с нашей моделью под глобальной смысловой связностью текста или его фрагмента будем понимать смысловую связь совокупности наименований понятий текста или его фрагмента, расположенных в определённом порядке. Под локальной смысловой связностью текста или его фрагмента будем понимать смысловую связь конкретного наименования понятия и его контекстного окружения.

Преобразование текстового представления в его формализованное смысловое представление дает возможность сопоставления текстов по их смысловому содержанию[12-13,15]. Такое сопоставление смыслового содержания текстов, обеспечивающее выявление близких по смыслу фрагментов текстов, на наш взгляд, должно удовлетворять следующим условиям:

В двух текстах должна быть пересекающаяся совокупность наименований понятий. Число понятий этой совокупности должно быть равно или превышать число наименований понятий, входящих в состав единичного высказывания.

В двух таких текстах должны быть фрагменты, в которых концентрация пересекающихся наименований понятий превышает пороговое значение. Эти фрагменты должны иметь соизмеримые размеры.

Эти фрагменты текстов должны быть сходными по составу наименований понятий и порядку их следования.

Определение схожего порядка следования наименований понятий в тексте или его фрагменте базируется на предположении, что смысл наименований понятий в значительной степени определяется их контекстным окружением [24-26]. В нашей модели смысл текста определяется как смысловое содержание совокупности взаимосвязанных наименований понятий, расположенных в нем в определенном порядке. Идентичные по смыслу тексты или их фрагменты должны удовлетворять условиям локальной и глобальной смысловой схожести. Локальная смысловая схожесть (ЛСС) наименований понятий текста определяется как сходство контекстного окружения идентичных наименований понятий в двух текстах или их фрагментах. Глобальная смысловая схожесть (ГСС) текстов или их фрагментов определяется как сходство состава идентичных наименований понятий и порядка их следования в текстах или их фрагментах. Каждое понятие этого фрагмента также должно удовлетворять условию локальной смысловой схожести.

Предлагаемая модель позволяет выявить близкие по тематике тексты или их фрагменты, после чего они, при необходимости, могут проверяться на смысловую идентичность.

2.2 Алгоритм выявления заимствований в текстах разноязычных документов

В результате проведенных исследований был разработан алгоритм выявления заимствований в текстах разноязычных документов. Необходимым условием для реализации этого алгоритма является использование многоязычного словаря унифицированных формализованных представлений наименований понятий. На данный момент в этом словаре содержатся слова и словосочетания на

русском и английском языках (общий объем словаря 3.5 млн. наименований понятий). Фрагмент многоязычного словаря унифицированных формализованных представлений наименований понятий приведен в таблице 1.

Таблица 1 Фрагмент многоязычного словаря унифицированных формализованных представлений наименований понятий

№ n/n	Основное значение в словаре	Синонимы	Эквиваленты на другом языке (английский)
…	…	…	…
816437	нефтехранилище	Нефтесклад / хранилище	oil reservoir / oil storage / petroleum storage / tank farm
816438	нефть	Каустобиолит / петролеум / черный золото	mineral oil / naphtha / oil / petrol / petroleum / rock-oil
816439	нефтяник	нефтедобытчик	Oilman / oil-industry worker
…	…	…	…

Также для работы этого алгоритма необходимы процедуры обработки текста для поддерживаемых языков. На данный момент используются процедуры для обработки текстов на русском и английском языках.

Далее приведем порядок выполнения алгоритма выявления заимствований в текстах разноязычных документов.

Шаг 1. Определяется язык анализируемого текста.

Шаг 2. Выявляется совокупность значимых наименований понятий с указанием местоположений этих понятий в тексте.

Шаг 3. Каждое наименование понятия с помощью процедуры автоматической пословной нормализации и словаря унифицированного формализованного представления наименований понятий приводится к унифицированной форме и ему присваивается номер из многоязычного словаря унифицированных формализованных представлений наименований понятий.

Шаг 4. Производится поиск совпадающих номеров наименований понятий в массиве формализованных представлений документов.

Шаг 5. Для рассматриваемого документа устанавливается перечень документов (документы могут быть на любом из поддерживаемых языков) близких ему по смысловому содержанию.

Шаг 6. Для пары документов - рассматриваемого документа и каждого из документов, найденных в п. 5, устанавливаются пары наиболее близких по смысловому содержанию фрагментов анализируемых текстов.

Шаг 7. Для каждой установленной в п.5 пары близких по смыслу фрагментов текстов определяется локальная смысловая схожесть всех наименований понятий этих фрагментов.

Шаг 8. Выбираются последовательности наименований понятий, имеющих значения локальной смысловой схожести выше заданного порога. Для каждой такой последовательности наименований понятий обоих текстов вычисляется степень их глобальной смысловой схожести.

2.3 Модель процесса выявления заимствований в текстах разноязычных документов

Модель для представления смыслового содержания текста в случае работы с разноязычными документами будет незначительно отличаться от использованной в предыдущих работах[19-23].

$КОДКО$ – концептуальный образ документа, дополненный контекстным окружением наименований понятий.

$КОДКО = \{НП_i, K_i \mid i \in [1, n_{НП}]\}$,

где $НП_i = (ННПС_i, Адр_i, \text{ОСРНП}_i, ЯНП_i)$;

$НП_i$ – информация об i-ом наименовании понятия;

$ННПС_i$ – номер наименования понятия в словаре многоязычном словаре унифицированных формализованных представлений наименований понятий;

$Адр_i$ – адреса вхождений наименования понятия в тексте;

ОСРНП_i – символ обобщенной синтаксической роли i-ого наименования понятия;

$ЯНП_i$ – язык i-ого наименования понятия;

$n_{НП}$ – количество наименований понятий;

K_i - множество контекстов i – ого наименования понятия, контексты описываются похожим образом:

$K_i = \{НПК_{ik} \mid k \in [1, n_{НПК_i}]\}$;

$НПК_{ik} = (ННПС_{ik}, Адр_{ik}, \text{ОСРНП}_{ik}, \text{КЗК}_{ik})$;

КЗК_{ik} – коэффициент значимости контекста;

Одним из важнейших этапов процесса выявления заимствований является вычисление мер выполнения условия локального и глобального смыслового сходства. Значение меры $м_{ik}$ выполнения условия локального смыслового сходства для каждого наименования понятия из КОДКО сравниваемых документов (в случае $м_{ik} = 0$ данное условие – не выполнено, при $м_{ik} > 0$ – выполнено частично, а при $м_{ik} = 1$ – выполнено полностью) вычисляется следующим образом:

Если $снп(\text{НП}_{pi}, \text{НП}_{jk}) = 0$, то $м_{ik} = 0$, иначе

$$м_{ik} = \frac{\text{снп}(\text{НП}_{pi}, \text{НП}_{jk})}{3} + \frac{2\text{ско}(\text{К}_{pil}, \text{К}_{jk\text{m}})}{3}$$

ско() – функция сравнения контекстного окружения наименований понятий;

$$\text{ско}(\text{К}_\text{a}, \text{К}_\text{b}) = \begin{cases} 1 & ,\text{фвзбк}(\text{К}_\text{a}, \text{К}_\text{b}) > 1 \\ \text{фвзбк}(\text{К}_\text{a}, \text{К}_\text{b}) & ,\text{фвзбк}(\text{К}_\text{a}, \text{К}_\text{b}) < 1 \end{cases}$$

ско() – функция вычисления значения близости контекстов;

$$\text{фвзбк}(\text{К}_\text{a}, \text{К}_\text{b}) = \frac{\sum_{c=0}^{n_{НПК_a}} \sum_{d=0}^{n_{НПК_b}} \text{фвппэ}(НПК_{ac}, НПК_{bd})}{4k_к}$$

фвппэ() – функция вычисления параметра похожести элементов контекстного окружения;

$k_к$ - размер контекста наименования понятия.

$\text{снп}(\text{НП}_{\text{pi}}, \text{НП}_{jk})$ – функция определения эквивалентности наименований понятий, причем $\text{снп}(\text{НП}_{\text{pi}}, \text{НП}_{jk}) \in \{0,1\}$, НП_{pi} – i-ый элемент формализованного смыслового описания рассматриваемого документа, НП_{jk} – k-ый элемент формализованного смыслового описания j-ого документа контрольного массива.

Условием глобального смыслового сходства является сходство порядка следования наименований понятий, но, поскольку порядок следования наименований понятий учтен при подсчете коэффициентов $м_{ik}$, с точностью до перестановок слов и словосочетаний, которые возможны в идентичных по смыслу текстах на одном языке или при переводе с одного языка на другой.

Для проверки выполнения условия глобального смыслового сходства необходимо произвести поиск последовательностей наименований понятий, у которых значения локальной смысловой схожести $м_{ik}$ выше некого заданного порога k_{ncx}. Мера выполнения условия глобального смыслового сходства вычисляется как среднее значение характеристик выполнения условия локального смыслового сходства содержащихся в этих последовательностях наименований понятий. Эта величина и будет являться коэффициентом смыслового сходства фрагментов текстов:

$$k_{cx} = \frac{\sum_{i=0}^{n_{НП_р}} \max_k(м_{ik})}{n_{НП_р}}$$

$\max_k(м_{ik})$ – максимальное значение $м_{ik}$,

при $k \in [1, n_{НП_j}]$; $n_{НП_р}$ – число элементов в КОДКО рассматриваемого документа; $n_{НП_j}$ – число элементов в КОДКО j-ого документа многоязычного контрольного массива.

3 Эксперимент выявления заимствований в текстах разноязычных документов

Для проверки работоспособности метода и возможности его использования в технологическом процессе выявления заимствований было принято решение провести небольшой эксперимент и посчитать показатели эффективности метода (полнота, точность и F1-мера). Для этого была собрана коллекция из 150 параллельных текстов (английский текст и его аутентичный перевод) по общественно-политической тематике. В процессе эксперимента русскоязычные тексты делились на предложения, для каждого из предложений определялись наиболее близкие по смысловому содержанию предложения англоязычных текстов. Пример установления смысловой близости двух разноязычных текстов приведен в таблице 2.

Таблица 2 Фрагменты параллельных текстов

Текст на русском языке	Текст на английском языке
….. Российские лидеры, конечно, беспокоятся о ценах на нефть, и для этого есть серьезная причина. Из-за падения цен на нефть падает стоимость рубля, сильно зависящая от этого показателя. Экспорт нефти важен для федерального бюджета и баланса внешней торговли России. Действительно, когда месячный курс цен на нефть марки Brent подскочил до 125 долларов за баррель в марте 2012 года, стоимость рубля приближалась к своему пику, около 29 рублей за один доллар. Когда цены на нефть упали до 30,70 доллара за баррель в январе 2016 года, стоимость рубля упала до 80 рублей за доллар. …..	….. Russia's leaders certainly do care about oil prices, and with good reason. Plunging oil prices decrease the ruble's value, which closely follows oil prices. Oil exports are important to Russia's federal budget and to its overall balance of trade. Indeed, when monthly average Brent oil prices peaked at about $125 per barrel in March 2012, the ruble was close to its own peak, at approximately twenty-nine rubles to every U.S. dollar. When Brent prices fell to $30.70 per barrel in January 2016, the ruble had fallen to about eighty rubles to the dollar. …..

Информация о текстах, участвующих в эксперименте, приведена в таблице 3.

Таблица 3 Информация о параллельных текстах

	Тексты на русском языке	Тексты на английском языке
Количество текстов	150	150
Количество предложений	6021	6021
Количество слов	157231	154863

Информация о результатах эксперимента приведена в таблице 4.

Таблица 4 Значения показателей эффективности метода

Полнота	Точность	F1 – мера
0.71	0.99	0.83

4 Заключение

В данной статье был предложен метод выявления заимствований в текстах разноязычных документов, базирующийся на семантико-синтаксическом и концептуальном анализе смысловой структуры разноязычных текстов. Разработанные на его основе алгоритмы были реализованы в виде экспериментального программного обеспечения, которое обеспечивает обработку текстов на двух языках (русском и английском). Эффективность предложенного метода была проверена на небольшой коллекции документов и показала удовлетворительные для первоначального этапа исследований результаты. Далее для улучшения качества работы метода необходимо будет провести дополнительную работу по модернизации алгоритмов и программного обеспечения, а также выполнить существенное пополнение словарей новой лексикой. Указанные мероприятия позволят значительно улучшить качество работы разработанных алгоритмов на текстах, относящихся к широкому спектру предметных областей. В настоящее время на рынке IT-услуг не существует промышленных программных средств, обеспечивающих сопоставление по их смысловому содержанию русскоязычных и англоязычных текстов. В связи с вышеизложенным нам представляется, что предлагаемый метод перспективен и кроме того он может иметь широкий спектр приложений.

Литература

[1] Potthast, Martin, Alberto Barron-Cedeno, Benno Stein, and Paolo Rosso. 2010. Cross-Language Plagiarism Detection. Language Resources and Evaluation, Special Issue on Plagiarism and Authorship Analysis, DOI: 10.1007/s10579-009-9114-z

[2] Alaa Zaid, Tiun Sabrina, Abdulameer Mohammedhasan Cross-language plagiarism of Arabic-English documents using linear logistic regression // Journal of Theoretical and Applied Information Technology, Vol. 83, No. 1, 10.01.2016, p. 20-33.

[3] Ceska Z., Toman, M, Jezek K. Multilingual Plagiarism Detection. // Artificial Intelligence: Methodology, Systems, and Applications, Proceedings of the 13th international conference on Artificial Intelligence: Methodology, Systems, and Applications, 2009, pp. 83-92.

[4] Chung-Hong Lee, Chih-Hong Wu, and Hsin-Chang Yang. 2008.A Platform Framework for Cross-lingual Text Relatedness Evaluation and Plagiarism Detection . The 3rd International Conference on Innovative Computing Information and Control (ICI-CIC'08).

[5] Mate Pataki A new approach for searching translated plagiarism . Proceedings of the 5th International Plagiarism Conference. Newcastle, UK, 2012.

[6] Ralf Steinberger Cross-lingual similarity calculation for plagiarism detection and more - Tools and resources. Keynotes for PAN 2012: Uncover-ing, Authorship, ad Social Software Misuse,2012.

[7] I.TRIFAN PLAGIARISM DETECTION IN A MULTILINGUAL ENVIRONMENT // Annals of DAAAM for 2011 & Proceedings of the 22nd International DAAAM Symposium, Volume 22, No. 1, ISSN 1726-9679, ISBN 978-3-901509-83-4, Editor B. Katalinic, Published by DAAAM International, Vienna, Austria, EU, 2011

[8] Tuomas Talvensaari Comparable Corpora in Cross-Language Information Retrieval (Academic Dissertation). Acta Electronica Universitatis Tamperensis 779, 2008.

[9] Diego Antonio Rodrıguez Torrejon, and Jose Manuel Martı Ramos Crosslingual CoReMo System. Notebook for PAN at CLEF 2011.

[10] Alexei Vinokourov, John Shawe-Taylor, and Nello Cristianini Inferring a Semantic Representation of Text via Cross-Language Correlation Analysis. Advances of Neural Information Processing Systems 15,2002.

[11] Philipp Cimiano, Antje Schultz, Sergey Sizov, Philipp Sorg, and Steffen Staab Explicit Versus Latent Concept Models for Cross-Language Information Retrieval. Proceedings of the Twenty-First International Joint Conference on Artificial Intelligence (IJCAI-09), 2009 .

[12] Кузнецов И.П. Механизмы обработки семантической информации. – М.: Наука, 1978. – 175 с.

[13] Осипов Г.С. Приобретение знаний интеллектуальными системами: Основы теории и технологии. – М.: Наука. Физматлит, 1997. – 112 с.

[14] Белоногов Г.Г. Теоретические проблемы информатики, Том 2. Семантические проблемы информатики. Под общей редакцией К.И. Курбакова. – М.: РЭА им. Г.В. Плеханова, 2008. – 342 с.

[15] Васильев В.Г., Кривенко М.П. Методы автоматизированной обработки текстов. – М.: ИПИ РАН, 2008. – 301 с.

[16] Лукашевич Н.В. Тезаурусы в задачах информационного поиска. М:, Изд. Моск. ун-та, 2011 г.- 508 с.

[17] Б. В. Добров, Н. В. Лукашевич Лингвистическая онтология по естественным наукам и

технологиям для приложений в сфере информационного поиска Учён. зап. Казан. гос. ун-та. Сер. Физ.-матем. науки, 149:2 (2007), 49–72

[18] Соссюр Фердинанд де. Курс общей лингвистики. – М.: Прогресс, 1977. – 370 с.

[19] Борзых А.И., Брагина Г.А., Хорошилов А.А. Методы автоматической кластеризации документов в хранилищах научно-технической информации для решения задачи поиска плагиата в текстах документов // Информатизация и связь. – 2012. – Вып. 8.

[20] Захаров В.Н., Хорошилов А.А. Автоматическая оценка подобия тематического содержания текстов на основе сравнения их формализованных смысловых описаний // Труды XIV-ой Всерос. науч. конф. «Электронные библио-теки: перспективные методы и технологии, электронные коллекции» – RCDL'2012, г. Переславль-Залесский, Россия, 15 – 18 октября 2012 г.

[21] Захаров В.Н., Хорошилов А.А. Методы решения задачи автоматического выявления заимствований в структурированных научно-технических документах на основе их семантического анализа // Труды XV-ой Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2013, г. Ярославль, 14 – 17 октября 2013 года.

[22] Хорошилов А.А. Методы выявления имплицитно выраженных заимствований в научно-технических текстах на основе их концептуального анализа // Труды XVII Международ-ной конференции «Аналитика и управление данными в областях с интенсивным использованием данных» DAMDID/RCDL'2015, Обнинск, 13 – 16 октября 2015 года. С. 471-477.

[23] Хорошилов А.А. Методы, модели, алгоритмы и экспериментальное программное обеспечение автоматического выявления неявно выраженных заимствований в научно-технических текстах.: дис. … канд. техн. наук: 05.13.17: защищена 09.12.15 – М.: 2015. – 159 с.

[24] Мельчук И.А. Опыт теории лингвистических моделей «Смысл ⇔ Текст». – М.: 1974 (2-е изд., 1999).

[25] Мельчук И.А. Русский язык в модели «Смысл ⇔ Текст». – Москва – Вена, 1995.

[26] Апресян Ю.Д., Богуславский И.М., Иомдин Л.Л. и др. Лингвистическое обеспечение системы ЭТАП-2. – М.: Наука, 1989.

[27] Белоногов Г.Г., Быстров И.И. и др. Автоматический концептуальный анализ текстов. // Научно-техническая информация. Сер. 2. – М.: ВИНИТИ, 2002. – № 10.

[28] Звегинцев В.А. Предложение и его отношение к языку и речи. – М.: Изд-во Московского университета, 1976.

A method of automatic plagiarism detection in multilingual documents

Victor N. Zakharov, Alexcandr A. Khoroshilov
Alexey A. Khoroshilov

The paper presents the method of automatic plagiarism detection in multilingual documents on the base of comparison of their formalized representations. In solving this problem, we developed a model of the semantic structure of texts. To detect plagiarism, we developed an algorithm for detection of similar semantic fragments in multilingual texts. The main advantage of this method is that it makes it possible to detect not only minor changes in the structure or lexical structure of the text, but also more complicated cases in the plagiarism.