

G.T.Yerkebulan, V.P.Kulikova.

Manash Kozybayev North Kazakhstan State University,

Petropavlovsk, Kazakhstan

E-mail: erkgulnur@mail.ru

COMPARATIVE ANALYSIS OF THE DETECTION SYSTEMS OF CROSS-LANGUAGE (TRANSLATIONAL) PLAGIARISM.

Abstract. The article deals with the problem of detection of cross-language or translated plagiarism in texts, describes the basic algorithms of finding cross-language borrowings. The analysis of plagiarism search systems in translated texts, their features and the main stages of the plagiarism detection process is also carried out. The article describes the algorithmic difficulties of the detection of translational plagiarism while using the automatic machine translation of texts from English into Russian. The specificity of the translated texts is analyzed, it consists in the discrepancy between the original text and the translated text due to the need of using different translation correspondences and equivalents, these results in rising complication of the process of identifying cross-language plagiarism in the translated texts. The methods used by the system "Antiplagiat" in the analysis of translated Russian texts for the purpose of detection of translated borrowings in them are considered. The results of monitoring and expert analysis of scientific articles included in the Russian index of scientific citation of publications and containing translated plagiarism are presented. A comparative analysis of the most popular programs of checking texts for plagiarism in terms of their functionality for recognition of translated plagiarism is carried out. The scheme of the algorithm actions of plagiarism detection systems of translation plagiarism is described.

Keywords: plagiarism, cross-language plagiarism, translated borrowings, plagiarism detection systems, anti-plagiarism programs.

УДК 00.004.9

Г.Т. Еркебулан, В.П. Куликова

Северо-Казахстанский государственный университет им. М Козыбаева

Петропавловск, Республика Казахстан

E-mail: erkgulnur@mail.ru

СРАВНИТЕЛЬНЫЙ АНАЛИЗ СИСТЕМ ОБНАРУЖЕНИЯ КРОСС-ЯЗЫКОВОГО (ПЕРЕВОДНОГО) ПЛАГИАТА

Аннотация. В статье рассматривается проблема обнаружения кросс-языкового или переводного плагиата в текстах, описываются основные алгоритмы поиска кросс-языковых заимствований. Также проводится анализ систем поиска плагиата в переводных текстах, их особенности и основные этапы процесса обнаружения плагиата. Описываются алгоритмические сложности обнаружения переводного плагиата в ходе использования

автоматизированного машинного перевода текстов с английского языка на русский. Анализируется специфика переводных текстов, которая состоит в несоответствии текста оригинала и текста перевода ввиду необходимости применения различных переводческих соответствий и эквивалентов, что в результате затрудняет процесс выявления кросс-языкового плагиата в переводных текстах. Рассматриваются методы, используемые системой «Антиплагиат» при анализе переводных русскоязычных текстов с целью обнаружения в них переводных заимствований. Представлены результаты мониторинга и экспертного анализа научных статей, входящих в Российский индекс научного цитирования изданий и содержащих переводной плагиат. Проведен сравнительный анализ наиболее популярных на сегодня программ проверки текстов на плагиат с точки зрения их функциональных возможностей по распознаванию переводного плагиата. Описана схема алгоритма действий системы обнаружения переводного плагиата.

Ключевые слова: плагиат, кросс-языковой плагиат, переводные заимствования, системы обнаружения плагиата, программы антиплагиат.

1. Введение. Проблема выявления оригинальности текста, особенно переводного, является актуальной в условиях современного информационного общества. Обычно при рассмотрении данной проблемы имеется в виду кросс-языковой плагиат, который связан с нарушением авторских прав.

Возможность обнаружения сходства текстовых документов, оригинального текста и переводного, позволяет повысить качество работы различных поисковых систем в Интернете, уменьшить заспамленность информационных текстов и порталов, систематизировать тексты в соответствии с их содержанием, а также выявить случаи нарушения авторских прав [1, с. 20-32].

Задача поиска похожих текстов является алгоритмически сложной ввиду значительных объемов обрабатываемых данных. В то же время алгоритмы, которые эффективны для решения одной задачи, могут оказаться неприменимыми или показывать неудовлетворительный результат при других задачах. Таким образом, к примеру, в юридической лингвистике исследуются стилистические и лексические особенности текстов, специфика их структуры, наличие оценочной лексики, категоричность и форма высказываний. Очевидно, что подобный подход окажется непродуктивным и, по сути, бесполезным при анализе переводных текстов. В связи с чем важно четко обозначить и разграничить понятия похожих текстов и кросс-языкового плагиата [2, с. 503-511].

Плагиатом считается умышленное присвоение и использование в собственных целях чужих идей, произведений, изобретений. Плагиат выражается в публикации чужих произведений от своего имени или без указания ссылок на заимствованные фрагменты [3, с. 93-114].

Плагиат обычно рассматривается с двух позиций: это может быть дословное копирование текста или копирование с элементами модификации. Именно ко второму случаю относятся переводные тексты, в которых может быть обнаружен кросс-языковой плагиат. Помимо этого, текст может быть скопирован полностью или же частично.

Под кросс-языковым или переводным плагиатом понимается полностью скопированный и переведенный на другой язык текст, т.е. в этом случае перевод выступает способом модификации текста [4, с. 33-40].

2. Методы. Существует различные методы нахождения плагиата в переводных текстах. Чаще всего процесс поиска происходит в соответствии со следующим алгоритмом:

- Эвристический поиск. В данном случае происходит отбор из корпуса документов, которые содержат фрагменты текста, похожие на некоторые части текста, который проверяется на плагиат. На этом этапе применяются различные алгоритмы выявления похожих текстов, а также зачастую необходимо использовать переход между языками документов – перевод ключевых слов, определение темы и жанра и др. [5]

- Подробный анализ. Осуществляется анализ степени сходства отобранных документов с проверяемым, и при обнаружении высокой степени сходства предполагается случай плагиата. Во многих системах антиплагиата этот шаг используется без эвристического поиска: сравнение идет с каждым из текстов корпуса.

- Пост-обработка. Полученные результаты анализа текста на плагиат проверяются с целью исключения ложных обнаружений заимствования (например, случаев, когда правильно оформленная цитата принимается за плагиат) [6].

Отметим, что при использовании машинного перевода иноязычных текстов с помощью различных онлайн переводчиков (например, Яндекс или Google) на русский язык и последующей проверке переводного текста системами антиплагиат (Антиплагиат.ру, E-txt, Advego, Text.ru) кросс-языковой плагиат не всегда обнаруживается системами. Это связано в первую очередь с тем, что проверки проводятся по русскоязычным коллекциям, а также с тем, что алгоритм поиска заимствований в указанных программах не подходит для проверки данного типа плагиата в анализируемых текстах. В результате переведенный текст, который является кросс-языковым плагиатом, определяется программой как оригинальная авторская работа [7, с. 121-126].

Кросс-языковой плагиат в русскоязычных текстах обычно возникает при переводе текстов с английского языка, поскольку именно он является международным общепринятым языком, применяемым на большинстве научных конференций и в большинстве научных изданий [8].

Для обнаружения плагиата сначала логично сделать перевод русскоязычного текста на английский язык, что можно осуществить также при помощи онлайн переводчика. Таким образом, поиск заимствований будет осуществляться по англоязычной коллекции. Но здесь появляется

новая трудность, которая состоит в том, что перевод характеризуется неоднозначностью и предполагает использование разнообразных переводческих эквивалентов и трансформаций, т.е. одна и та же мысль может передаваться посредством разных слов и словосочетаний, кроме того, при грамотном переводе изменяется порядок слов и структура текста и предложений. Ситуация усугубляется тем, что при автоматическом машинном переводе возникают дополнительные ошибки, недочеты и несоответствия. В результате в переведенном тексте при помощи традиционных алгоритмов невозможно корректно определить заимствования [9, с. 5-7].

Рассмотрим процесс обнаружения плагиата в переводных текстах на примере системы «Антиплагиат». Разработчиками программы был создан модуль выявления переводных заимствований для языковой пары «английский – русский». Данный метод дает возможность осуществлять в коллекциях поиск документов-кандидатов, откуда потенциально могли быть произведены заимствования, и сравнивать текст. Такая технология применима и для других языковых пар.

Главным методом поиска документов-кандидатов при дословном заимствовании текста выступает построение инвертированного индекса, где документ из коллекции представляется в виде набора перекрывающихся друг друга последовательностей слов определенной длины (n -грамм). Далее происходит упорядочивание документов-кандидатов на основании выбранной функции схожести совпавших последовательностей слов. В случае переводных заимствований задача обнаружения документов-кандидатов решается аналогичным методом. Основное отличие от классического способа обнаружения документов-кандидатов состоит в построении функции, которая ставит в соответствие каждому слову в рассматриваемой последовательности класс эквивалентности – метку кластера, объединяющего слова, которые близки по смыслу и встречаются в одинаковых контекстах. Модели на основе глубоких сетей обеспечивают представление слов в векторном пространстве, которое может использоваться с целью выявления семантической близости слов. Таким образом, кластеризация векторов выступает быстрым и эффективным методом получения классов слов [10, с. 1-2].

На следующем этапе осуществляется сравнение слов между собой. Для этого применяются рекуррентные сети глубокого обучения – особый тип искусственных нейронных сетей, которые дают возможность представить часть текста произвольной длины в векторном пространстве. С целью оптимизации параметров сети используются технологии частичного обучения (с небольшим объемом частично размеченных данных) и технологии обучения без учителя (unsupervised learning) – машинное обучение, где известны только описания множества объектов и требуется обнаружить внутренние зависимости, существующие между ними. Вопрос обработки естественного языка, таким образом, разрешается с меньшими

усилиями по поиску подходящих ресурсов, что актуально для многих языков, по которым имеется недостаточное количество материалов [11, с. 532-528].

Еще одной значимой особенностью рекуррентных сетей представляется возможность их использования как порождающих, т.е. моделирующих вероятностные характеристики естественного языка, моделей. Такая технология дает возможность генерировать тексты на основании изученных языковых структур, что, в свою очередь, позволяет использовать их не только при выявлении семантической близости слов, словосочетаний и выражений, но и при перефразировании. Оптимальная архитектура для каждого вида нейронных сетей подбирается на основании значений конкретных целевых метрик оценки качества: точности, полноты и F-меры, которая учитывает при оценке и точность, и полноту [12].

3. Результаты. Действие программы тестировалось на научных статьях, которые входят в Российский индекс научного цитирования (РИНЦ) изданий, предоставленных научной электронной библиотекой eLibrary.ru [13, с. 61-68]. Мониторингу подверглось примерно 2,5 млн. русскоязычных научных статей, проверка производилась по англоязычным коллекциям. Итак, в результате мониторинга было выявлено более 20 тыс. русскоязычных научных статей, в которых содержался в больших объемах кросс-языковой плагиат. Около трети таких статей прошли дополнительный экспертный анализ, результаты которого показаны в таблице 1.

Таблица 1 – Результаты экспертного анализа статей, содержащих кросс-языковой плагиат [12]

Результаты анализа статей с переводными заимствованиями (кросс-языковым плагиатом)	
Тип срабатывания системы	Количество случаев
Заимствования, в том числе:	2627
– переводные заимствования (текст переведен с английского языка на русский и выдан за оригинальный)	921
– другие заимствования (из русскоязычных ресурсов)	1706
Легальные заимствования, в том числе:	2355
– двуязычные статьи (статьи одного автора на двух разных языках)	788
– цитирование законодательных актов (использование в тексте формулировок нормативных актов)	1567
Самоцитирование (цитирование автором своей оригинальной англоязычной работы)	660
Ошибочные срабатывания	507
Другое (проверяемые статьи затруднительно	1540

однозначно отнести к определенной категории)	
Итого	7689

Таким образом, мы видим, что часть проанализированных статей представляет собой легальные заимствования. Данные статьи являются переводными работами одних и тех же авторов или соавторов, выполненными на разных языках. Но важно отметить, что среди статей значительная часть определена как некорректные переводные заимствования, т.е. кросс-языковой плагиат.

В данном случае можно считать, что в электронной библиотеке eLibrary.ru как минимум 1% русскоязычных научных статей содержат в различном объеме переводной плагиат. Необходимо учитывать, что этот процент значительно занижен, поскольку проверка ограничивалась размером англоязычной коллекции документов. Следовательно, при индексировании большего количества текстов на английском языке можно обнаружить большее количество текстов с переводным плагиатом [14].

Сравнивая существующие системы поиска переводного плагиата, можно отметить, что наиболее эффективной программой при решении данного вопроса на сегодняшний день является полная версия «Антиплагиат» со специально разработанным модулем поиска переводных заимствований. При этом неполная бесплатная версия программы не распознает плагиат в переводных текстах.

Система «Advego Plagiatyus»: программа имеет функции быстрой и глубокой проверки текстов, а также осуществляет проверку на плагиат не только по словам, но и по фразам, т.е. на смысловом уровне, но при проверке переводных текстов также не распознает плагиат.

Программа «Text.ru» проводит полный SEO-анализ текста в режиме онлайн и очень полезна при анализе текстов на наличие рерайта, т.е. перефразирования, однако не способна распознавать переводной плагиат и расценивает переводные тексты с английского на русский язык как полностью оригинальные.

Система обнаружения плагиата «E-txt» при проверке переводных текстов выдает множество ошибок при чтении файлов и показывает очень высокий уровень оригинальности текста, т.е. также не способна распознать переводной плагиат [15].

4. Обсуждение. Несанкционированное использование оригинального контента, или плагиат, является серьезной проблемой. Плагиат в переводных текстах обнаружить достаточно сложно. С этой задачей большинство программ проверки текстов на плагиат не справляется и показывает высокий уровень оригинальности текста при анализе переводных документов.

Однако, подключение разработанного модуля поиска переводных заимствований в полной версии системы «Антиплагиат» позволяет эффективно распознавать кросс-языковой плагиат в различных текстах.

В результате анализа систем обнаружения переводного плагиата можно обозначить следующую схему алгоритма действий для поиска заимствований с английского на русский язык:

- проверяемый русскоязычный текст поступает на вход;
- производится машинный перевод данного текста с русского языка на английский;
- в проиндексированной коллекции текстов на английском языке осуществляется поиск документов-кандидатов в источники заимствования;
- происходит сопоставление каждого обнаруженного документ-кандидата с английской версией проверяемого текста, определяются границы заимствованных фрагментов;
- границы заимствованных фрагментов переносятся в русскоязычную версию текста;
- по окончании процесса проверки на переводной плагиат составляется отчет о результатах.

Список литературы

[1] Zanettin F. Corpus methods for descriptive translation studies // *Procedia-Social and Behavioral Sciences*. – 2017. – Iss. 95. – P. 20-32.

[2] Identification of translationese: A machine learning approach // *Proceedings of the 11th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2010)*. – Springer Heidelberg, March 2015. – P. 503-511.

[3] Lapshinova-Koltunski E. Variation in translation: Evidence from corpora // *New directions in corpus-based translation studies* / C. Fantinuoli, F. Zanettin. – Berlin: Language Science Press, 2015. – P. 93-114.

[4] Kunilovskaya M., Kutuzov A. A quantitative study of translational Russian (based on a translational learner corpus) // *Corpus Linguistics 2015: Proceedings of 7th International Conference*. – St. Petersburg, 2015. – P. 33-40.

[5] Ilisei I. A Machine Learning Approach to the Identification of Translational Language: An Inquiry into Translationese Learning Models. – Wolverhampton: University of Wolverhampton, 2015.

[6] Астапова О.П. Исследование и разработка методов поиска плагиата в многоязычных корпусах текстов. [Электронный ресурс]. – Режим доступа: <http://seminar.at.ispras.ru/wp-content/uploads/2012/07/Astapova-thesis1.pdf> (Дата обращения: 16.04.2019).

[7] Шарапов Р.В., Шарапова Е.В. Система проверки текстов на заимствование из других источников // *Труды 15-й Всероссийской научной конференции «Электронные библиотеки: Перспективные методы и технологии, электронные коллекции»*. – Воронеж, 2014. – С. 121-126.

[8] Мазов Н.А., Гуреев В.Н. Результаты исследований по выявлению переводного плагиата с использованием библиометрических баз данных // *Научные и технические библиотеки*. – 2017. – № 12. [Электронный ресурс]. – Режим доступа: <https://ntb.gpntb.ru/jour/article/view/257#> (Дата обращения: 17.04.2019).

[9] Аббаров Р.Д. Разработка методов поиска плагиата // Молодой ученый. – 2016. – №12.4. – С. 5-7.

[10] Аббаров Р.Д., Худайберганова М.М. Как работает антиплагиат? Принцип работы антиплагиата // Молодой ученый. – 2016. – №29.3. – С. 1-2.

[11] Nisioi S., Dinu L.P. A clustering approach for translationese identification // Proceedings of RANLP. – 2018. – P. 532-538.

[12] Чехович Ю., Кузнецова Р., Бахтеев О. Плагиат в научных статьях: трудности обнаружения перевода. [Электронный ресурс]. – Режим доступа: <http://www.unkniga.ru/innovation/tehnology/7886-plagiat-v-nauchnyh-statjah0trudnosti-obnaruzheniya-perevoda.html> (Дата обращения: 17.04.2019).

[13] Никитов А.В., Орчаков О.А., Чехович Ю.В. Плагиат в работах студентов и аспирантов: проблема и методы противодействия // Университетское управление: практика и анализ. – 2014. – № 5. – С. 61-68.

[14] Kunilovskaya M., Morgoun N., Pariy A. Learner vs. Professional Translations into Russian: Lexical Profiles // Translation and Interpreting. – 2017. – Vol. 10, No. 2.

[15] Звонникова Е.Д. Сравнение информационных систем автоматической проверки оригинальности текстов // Научно-методический электронный журнал «Концепт». – 2015. – № 8 (август). – С. 126-130. [Электронный ресурс]. – Режим доступа: <http://e-koncept.ru/2015/15282.htm> (Дата обращения: 15.04.2019).

References

[1] Zanettin F. Corpus methods for descriptive translation studies // Procedia-Social and Behavioral Sciences. – 2017. – Iss. 95. – P. 20-32.

[2] Identification of translationese: A machine learning approach // Proceedings of the 11th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2010). – Springer Heidelberg, March 2015. – P. 503-511.

[3] Lapshinova-Koltunski E. Variation in translation: Evidence from corpora // New directions in corpus-based translation studies / C. Fantinuoli, F. Zanettin. – Berlin: Language Science Press, 2015. – P. 93-114.

[4] Kunilovskaya M., Kutuzov A. A quantitative study of translational Russian (based on a translational learner corpus) // Corpus Linguistics 2015: Proceedings of 7th International Conference. – St. Petersburg, 2015. – P. 33-40.

[5] Ilisei I. A Machine Learning Approach to the Identification of Translational Language: An Inquiry into Translationese Learning Models. – Wolverhampton: University of Wolverhampton, 2015.

[6] Astapova O.P. Issledovanie i razrabotka metodov poiska plagiata v mnogoyazyichnyih korpusah tekstov. [Elektronnyiy resurs]. – Rezhim dostupa: <http://seminar.at.ispras.ru/wp-content/uploads/2012/07/Astapova-thesis1.pdf> (Data obrascheniya: 16.04.2019).

[7] Sharapov R.V., Sharapova E.V. Sistema proverki tekstov na zaimstvovanie iz drugih istochnikov // Trudy 15-y Vserossiyskoy nauchnoy konferentsii «Elektronnyie biblioteki: Perspektivnyie metody i tehnologii, elektronnyie kollektsii». – Voronezh, 2014. – S. 121-126.

[8] Mazov N.A., Gureev V.N. Rezultaty issledovaniy po vyyavleniyu perevodnogo plagiata s ispolzovaniem bibliometricheskikh baz danniyh // Nauchnyie i tehicheskie biblioteki. – 2017. – # 12. [Elektronnyiy resurs]. – Rezhim dostupa: <https://ntb.gpntb.ru/jour/article/view/257#> (Data obrascheniya: 17.04.2019).

[9] Abrarov R.D. Razrabotka metodov poiska plagiata // Molodoy ucheniy. – 2016. – #12.4. – S. 5-7.

[10] Abrarov R.D., Hudayberganova M.M. Kak rabotaet antiplagiat? Printsip raboty antiplagiata // Molodoy ucheniy. – 2016. – #29.3. – S. 1-2.

[11] Nisioi S., Dinu L.P. A clustering approach for translationese identification // Proceedings of RANLP. – 2018. – P. 532-538.

[12] Chehovich Yu., Kuznetsova R., Bahteev O. Plagiat v nauchnyih statyah: trudnosti obnaruzheniya perevoda. [Elektronnyiy resurs]. – Rezhim dostupa: [http://www.unkniga.ru/innovation/tehnology/7886-plagiat-v-nauchnyh-statjah0trudnosti-obnaruzheniya-perevoda.html](http://www.unkniga.ru/innovation/tehnology/7886-plagiat-v-nauchnyh-statyah0trudnosti-obnaruzheniya-perevoda.html) (Data obrascheniya: 17.04.2019).

[13] Nikitov A.V., Orchakov O.A., Chehovich Yu.V. Plagiat v rabotah studentov i aspirantov: problema i metody protivodeystviya // Universitetskoe upravlenie: praktika i analiz. – 2014. – # 5. – S. 61-68.

[14] Kunilovskaya M., Morgoun N., Pariy A. Learner vs. Professional Translations into Russian: Lexical Profiles // Translation and Interpreting. – 2017. – Vol. 10, No. 2.

[15] Zvonnikova E.D. Sravnenie informatsionnyih sistem avtomaticheskoy proverki originalnosti tekstov // Nauchno-metodicheskiy elektronnyiy zhurnal «Kontsept». – 2015. – # 8 (avgust). – S. 126-130. [Elektronnyiy resurs]. – Rezhim dostupa: <http://e-koncept.ru/2015/15282.htm> (Data obrascheniya: 15.04.2019).

G.T.Yerkebulan, V.P.Kulikova.

Comparative analysis of the detection systems of cross-language (translational) plagiarism.

Abstract. The article deals with the problem of detection of cross-language or translated plagiarism in texts, describes the basic algorithms of finding cross-language borrowings. A comparative analysis of the most popular programs of checking texts for plagiarism in terms of their functionality for recognition of translated plagiarism is carried out. The scheme of the algorithm actions of plagiarism detection systems of translation plagiarism is described.

Keywords: plagiarism, cross-language plagiarism, translated borrowings, plagiarism detection systems, anti-plagiarism programs.

Г.Т. Еркебұлан, В.П. Куликова.

Тілдераралық (аудармалы) плагиатты анықтау жүйелерін салыстырмалы талдау.

Түйіндеме. Мақалада мәтіндерде тілдераралық немесе аудармалы плагиатты анықтаудың мәселелері қаралған, тілдераралық (рұқсатсыз) алып пайалануды іздеудің негізгі алгоритмдері баяндалған. Аудармалы плагиатты анықтаудың функционалдық мүмкіндіктері жағынан біүгінгі күні барынша танымал мәтіндерді плагиатқа тексеретін

бағдарламаларға салыстырмалы талдау жүргізілген. аудармалы плагиатты анықтау жүйесінің іс-әрекеттер алгоритмінің сызбасы сипатталған.

Түйінді сөздер: плагиат, тілдераралық плагиат, аудармалы (рұқсатсыз) пайдалану, плагиатты анықтау жүйесі, антиплагиат бағдарламасы.

Г.Т. Еркебулан, В.П. Куликова.

Сравнительный анализ систем обнаружения кросс-языкового (переводного) плагиата.

Аннотация. В статье рассматривается проблема обнаружения кросс-языкового или переводного плагиата в текстах, описываются основные алгоритмы поиска кросс-языковых заимствований. Проведен сравнительный анализ наиболее популярных на сегодня программ проверки текстов на плагиат с точки зрения их функциональных возможностей по распознаванию переводного плагиата. Описана схема алгоритма действий системы обнаружения переводного плагиата.

Ключевые слова: плагиат, кросс-языковой плагиат, переводные заимствования, системы обнаружения плагиата, программы антиплагиат.