УДК: 001.893:519.7:81`32

ОБ ОДНОМ МЕТОДЕ ДЕТЕКТИРОВАНИЯ ИСКУССТВЕННЫХ И НЕНАУЧНЫХ ТЕКСТОВ В ОБШИРНОЙ КОЛЛЕКЦИИ ДОКУМЕНТОВ

О.Ю. Бахтеев[1], М.В. Кузнецова[2], А.В. Романов[3], Ю.В. Чехович[4]

[1-4]Компания «Антиплагиат» (115093, г. Москва, ул. Большая Серпуховская, д. 44, офис 33)

[1]bakhteev@ap-team.ru, [2]kuznetsova@ap-team.ru,
[3]alexey.romanov@phystech.edu, [4]chehovich@antiplagiat.ru

Аннотация

Работа посвящена описанию метода детектирования искусственных и ненаучных текстов в коллекции научных статей. Предлагаемый метод основан на лексическом и морфологическом анализе проверяемого документа, позволяющем оценить вероятность его принадлежности к классу научных документов. Эксперименты подтверждают возможность практического применения метода.

Ключевые слова: *обработка естественного языка, классификация документов, анализ текстов, статистические языковые модели, детектирование искусственных текстов.*

ВВЕДЕНИЕ

Проблема детектирования искусственных текстов в последние годы вызывает растущий интерес по ряду объективных причин: с одной стороны, развитие технологий хранения и обработки цифровых документов позволяет осуществлять взаимодействие редакционных коллегий научных журналов, издательств, организаторов конференций с авторами в электронном формате. С другой стороны, наблюдается появление множества алгоритмов автоматической генерации текстов на естественном языке, внешне неотличимых от текстов, написанных человеком, но не несущих смысловой нагрузки и, тем более, не обладающих научной значимостью.

В данных условиях становится доступным для широких масс не требующий высоких затрат способ подготовки псевдонаучных документов, рецензирование которых экспертами невозможно без детального анализа. В то же время имеет

место явление, при котором ненаучный контент попадает в документы, поданные на рецензирование, вследствие случайных ошибок автора и технических сбоев программного обеспечения. Кроме того, часто возникают ситуации, когда агрегаторами, библиометрическими и поисковыми сервисами «по определению» относятся к научным все публикации в научном издании, в то время как среди них могут быть редакционные статьи, отчеты организаций, поздравления с юбилеями и другие подобные произведения.

В настоящей работе предложен метод автоматической фильтрации текстов без привлечения экспертов. Метод основан на использовании статистических языковых моделей, строящихся по корпусам предварительно подготовленных текстов. Данные модели позволяют оценить вероятность принадлежности текста к классу научных документов и на основании полученной оценки автоматически отфильтровывать ненаучные и искусственные тексты на начальном этапе анализа, до привлечения экспертов. Проведённые эксперименты на реальных и сгенерированных с помощью компьютера документах подтверждают практическую применимость метода.

Следует отметить, что авторы работы не претендуют на использование разработанного метода в качестве средства оценки научной значимости статьи или сравнительного анализа статей, а рассматривают исключительно внешние признаки исследуемых текстов.

1. ОБЗОР ЛИТЕРАТУРЫ

Задачи, связанные с выявлением текстов, сгенерированных с помощью компьютера, активно рассматриваются в научном сообществе в течение последнего десятилетия. Так, например, в [2] описан алгоритм, направленный на обнаружение текстов, созданных с помощью конкретного доступного инструмента – SciGen, широко используемого англоязычными авторами [3].

Метод, описанный в [4], предназначен, в первую очередь, для детектирования автоматически генерируемого веб-спама, но идеи, лежащие в его основе, могут быть использованы и при решении задачи обнаружения более широкого класса искусственных документов.

Предлагаемый в данной работе подход опирается на результаты работы [1] и адаптирует их к морфосинтаксическим особенностям русского языка. Лекси-

ческий и морфологический анализ научных, ненаучных и искусственных текстов позволяет эффективно определять принадлежность документа к одному из классов.

2. ЛЕКСИЧЕСКИЕ И МОРФОСИНТАКСИЧЕСКИЕ ОСОБЕННОСТИ НАУЧНОГО ТЕКСТА

С точки зрения анализа текста, научные документы обычно характеризуются различными лексическими особенностями:

- употребление специфической терминологии;
- использование речевых клише;
- употребление лексических единиц с абстрактным значением.

Обобщённость научной речи проявляется в особенностях употребления различных морфологических единиц, выборе конструкций и частотности их появления в тексте. Вместе с тем автоматические генераторы текстов, как правило, имеют ограниченный функционал в этом плане: зачастую в их основе лежит небольшое число шаблонов, по которым алгоритмически создаются фразы и предложения. В большинстве случаев детектирование подобных шаблонов ведёт к выделению сгенерированных текстов среди научных документов.

3. ОПИСАНИЕ ПРЕДЛАГАЕМОГО МЕТОДА

Статистические языковые модели (СЯЗ) – механизм, позволяющий оценивать частотность появления в тексте отдельных слов и N-грамм – последовательностей из N слов, идущих подряд. Благодаря таким моделям, построенным на одном из классов документов, можно оценивать вероятность принадлежности проверяемого документа к рассматриваемому классу как интегральную функцию оценок, полученных для отдельных слов и N-грамм.

Алгоритм, описываемый в данной работе, предполагает использование СЯЗ, которые строятся как на последовательностях слов (лексические СЯЗ), так и на последовательностях меток, приписываемым словам текста морфологическим анализатором (морфологические СЯЗ). Первые направлены на выявление случаев употребления ненаучной лексики или несочетаемых друг с другом терминов, а вторые – использования шаблонов, характерных для автоматических текстовых генераторов.

Комбинация моделей обоих типов позволяет выявлять как ненаучные, так и искусственные тексты путём классификации проверяемого документа как подозрительный в случае, если оценка вероятности его принадлежности к классу научных документов, получаемая с помощью СЯЗ, ниже предварительно устанавливаемого порога.

4. РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТОВ

Для получения оценок частотности слов и N-грамм, используемых при создании СЯЗ, были подготовлены корпуса:

- научных документов (10 тыс. статей, монографий, диссертаций и авторефератов, находящихся в открытом доступе);
- ненаучных документов (25 тыс. книг художественной и публицистической литературы);
- автоматически сгенерированных документов (10 тыс. документов, созданных с помощью генераторов различных типов [2], [4]).

Решалась задача классификации проверяемых текстов на 2 класса:

- научные документы;
- ненаучные и искусственные документы.

Отдельная выборка данных, представляющая собой часть совокупности подготовленных корпусов, использовалась для тестирования алгоритма и оценки его качества. Измерялись следующие показатели качества алгоритма классификации на данной выборке:

- точность (precision) – 99%;
- полнота (recall) – 84%;
- F-мера – 91%;

Разработанный алгоритм был применён для поиска искусственных и ненаучных текстов в коллекции документов научной электронной библиотеки eLIBRARY.RU, содержащей более 3,3 млн. документов. В результате работы алгоритма более 250 тыс. документов были помечены как подозрительные. Анализ ошибок выявил следующие классы документов, выделенных алгоритмом:

- тексты, содержащие типовые повторяющиеся упражнения и контрольные вопросы;
- тексты, содержащие таблицы и пояснения к таблицам;

- рефераты, списки примечаний и библиографии;
- поздравления с юбилеем, некрологи, итоги работы предприятий;
- рекламные тексты;
- прочие типы документов

Текстов, сгенерированных с помощью автоматических средств, в коллекции eLIBRARY.RU алгоритмом выявлено не было.

ЗАКЛЮЧЕНИЕ

В работе предложен метод автоматического обнаружения ненаучных и искусственных текстов в коллекции научных документов. Метод основан на использовании статистических языковых моделей, позволяющих оценить вероятность принадлежности проверяемого текста к классу научных документов. Для оценивания вероятности используются языковые модели, построенные на последовательностях слов и последовательностях меток, приписываемых этим словам морфологическим анализатором. Эксперименты, проведённые на реальных данных, показали практическую применимость метода.

Варианты дальнейших исследований по данной теме включают в себя повышение качества работы алгоритма на имеющихся данных, его адаптацию к другим классам искусственных текстов и разработку методов автоматического оценивания качества научного текста и его научной значимости.

СПИСОК ЛИТЕРАТУРЫ

1. Arase Y., Zhou M. Machine Translation Detection from Monolingual Web-Text // ACL (1). 2013. P. 1597–1607.

2. Labbé C., Labbé D. Duplicate and fake publications in the scientific literature: how many SCIgen papers in computer science? //Scientometrics. 2013. V. 94, No 1. P. 379–396.

3. Van Noorden R. Publishers withdraw more than 120 gibberish papers //Nature. 2014. V. 24.

4. Гречников Е. А. и др. Поиск неестественных текстов // Тр. XI Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции». Петрозаводск, 2009. С. 306–308.

A METHOD FOR DETECTING ARTIFICIAL AND NON-SCIENTIFIC TEXTS IN THE COLLECTION OF DOCUMENTS

O.Yu. Bakhteev[1], M.V. Kuznetsova[2], A.V. Romanov[3], Yu.V. Chekhovich[4]

[1-4] Antiplagiat Company (115093, Moscow, Bolshaya Serpuhovskaya, 44, office 33)

[1]bakhteev@ap-team.ru, [2]kuznetsova@ap-team.ru,

[3]alexey.romanov@phystech.edu, [4]chehovich@antiplagiat.ru

Abstract

In this paper, we propose a method of machine-generated and non-scientific text detection in a collection of scientific papers. The method is based on lexical and morphological analysis of the document examined with the help of language modeling. This technique enables estimation of probability that the text belongs to the class of scientific documents. Experimental evidence shows feasibility of the approach.

Keywords: *natural language processing, document classification, text mining, statistical language models, machine-generated text detection.*

REFERENCES

1. Arase Y., Zhou M. Machine Translation Detection from Monolingual Web-Text //ACL (1). 2013. P. 1597–1607.

2. Labbé C., Labbé D. Duplicate and fake publications in the scientific literature: how many SCIgen papers in computer science? //Scientometrics. 2013. V. 94, No 1. P. 379–396.

3. Van Noorden R. Publishers withdraw more than 120 gibberish papers //Nature. 2014. V. 24.

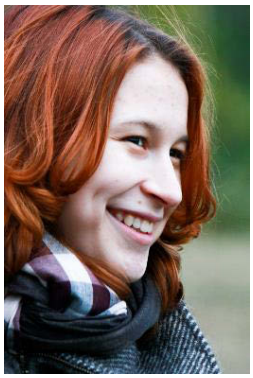
4. Grechnikov E.A. i dr. Poisk neestestvennykh tekstov // Tr. XI Vserossiyiskoy nauchnoi konferencii “Electronnye bibloitheki: perspernivnye methody b tekhnologii, electronnye kollekcii”, Petrozavodsk, 2009. S. 306–308.

СВЕДЕНИЯ ОБ АВТОРАХ

БАХТЕЕВ Олег Юрьевич – старший исследователь компании Антиплагиат

Oleg Yurievich BAKHTEEV – senior researcher, Antiplagiat Company

email: bakhteev@ap-team.ru



КУЗНЕЦОВА Маргарита Валерьевна – руководитель отдела исследований компании Антиплагиат

Margarita Valerievna KUZNETSOVA – head of research department, Antiplagiat Company

email: kuznetsova@ap-team.ru

РОМАНОВ Алексей Владимирович – ассистент, компания Abbyy

Alexey Vladimirovich ROMANOV – assistant, Abbyy Company

email: alexey.romanov@phystech.edu

ЧЕХОВИЧ Юрий Викторович – исполнительный директор компании Антиплагиат, кандидат физико-математических наук

Yury Viktorovich CHEKHOVICH – Chief Executive Officer, Antiplagiat Company, PhD (Mathematics)

email: chehovich@antiplagiat.ru

Материал поступил в редакцию 21 октября 2017 года