Борьба с синонимизацией в текстовых потоках данных

С.Н. Середа, Е.В. Шарапова

Муромский институт (филиал) ФГБОУВО «Владимирский государственный университет имени Александра Григорьевича и Николая Григорьевича Столетовых»
602264, г. Муром, Владимирской обл., ул. Орловская, 23
E-mail: mivlgu@mail.ru

*Работа посвящена вопросам борьбы с использованием синонимизации текстов для сокрытия заимствований. Изложен подход, предполагающий замену всех слов текста их наиболее весомыми синонимами. Вес синонимов определяется в соответствии с частотой их встречаемости в русском языке. Для слов, имеющих вес выше веса синонимов, замена не производится. Такой подход позволяет сравнивать измененные тексты с использованием метода шинглов, TF*IDF, сигнатур и т.д., корректно оценивать долю оригинальности текстовых документов.*
*The work is devoted to the issues of combating the use of synonymization of texts to hide borrowings. An approach is presented that assumes the replacement of all words of the text with their most significant synonyms. The weight of synonyms is determined in accordance with the frequency of their occurrence. For words with a weight higher than the weight of synonyms, no substitution is made. This approach allows to compare modified texts using the shingle method, TF * IDF, signatures, etc., correctly assess the share of originality of text documents.*

Введение

Синонимизация – замена слов в тексте синонимами (словами со схожим смыслом, но различным написанием) [1]. Основная цель синонимизации состоит в изменении текстового документа таким образом, чтобы повысить его уникальность скрыв тем самым факт заимствования [2]. Она состоит в замене некоторых слов текста на синонимы. Синонимизация реализуется как вручную, так и в автоматическом режиме с помощью различных онлайн сервисов и программ (синонимайзеров). В настоящее время существует достаточно большое количество синонимайзеров – Raskruty.ru, Usyn.ru, Seogenerator.ru, Seo-builder.ru, Sinoni.men, Sinonimov.ru, Rustxt.ru, Textrobot.ru, Online-sinonim.ru, Synonymizer.ru, Progaonline.com/synonymizer/, Fromtlt.ru/sinonim и т.д. Они в значительной степени отличаются используемыми базами синонимов и алгоритмами работы (в первую очередь степенью переработки текста и его «читаемостью»). При синонимизации часто всплывает такой негативный момент, как неудачный подбор синонимов, искажающий смысл оригинального текста [3].

Синонимизация текстов представляет собой явление сокрытия заимствований текстов, с которым необходимо бороться. Несмотря на многие попытки [4, 5, 6, 7] в настоящее время еще не найдены эффективные подходы, позволяющие нейтрализовать синонимизацию. Цель работы – рассмотреть вопросы борьбы с синонимизацией в текстовых документах.

Борьба с синонимизацией

Поставим каждому слову w_i в словаре W его вес f_i, подсчитанный на основе глобальной частоты его встречаемости в русскоязычных текстах.

Для каждого слова составляется список синонимов $w_i=\{s_{1,} s_{i2} \ldots s_{in}\}$ и их весов $f_i=\{ f_{i1,} f_{i2} \ldots f_{in}\}$. Список синонимов подбирается на основе базы синонимов SynMaster, данные о частоте слов по [8].

Тогда наиболее представительным синонимом будет слово с максимальным значением веса, т.е. $max(f_{i1,} f_{i2} \ldots f_{in}) \rightarrow s_i$. Если вес синонима s_i превышает вес слова w_i, то синоним принимается как кандидат на замену слова, в противном случае процедура поис-

ка синонимов прекращается. Далее процедура итерационно повторяется для вновь найденного синонима до тех пор, пока вес слова-синонима не станет максимальным. Процедура прекращается при весе вновь найденного синонима менее веса текущего слова кандидата s_i. На последнем шаге слово w_i заменяется найденным синонимом s_i. Надо заметить, что при весе слова больше весов всех кандидатов на синонимы, его замена синонимами не производится, а слово считается наиболее представительным.

Обработка текста заключается в замене всех содержащихся слов их наиболее весомыми («тяжелыми») синонимами. Конечно, преобразованный таким образом текст, становится очень далек от оригинала как по содержанию, так и по смыслу. Тем не менее, текстовые документы, предварительно прошедшие синонимазацию для сокрытия заимствований, после обработки описанным выше способом будут приведены примерно к тому же виду. Поэтому сравнение их содержания даст вполне приемлемые результаты.

С одной стороны, результат синонимизации в значительной степени зависит от используемой базы синонимов. По этой причине, процедура однократной обработки синонимов может не привести к получению текстов, близких по содержанию. Тем не менее, осуществление итерационного поиска синонимов дает возможность приблизиться к этому.

Исследования показали, что часто в качестве синонимов могут встречаться местоимения и другие часто употребляемые слова. Рассмотрим, например, список синонимов для слова «автор»:

автор | *творец* | *создатель* | *писатель* | *сочинитель* | *литератор* | *композитор* | *виновник* | *либреттист* | *составитель* | *компилятор* | *пишущий эти строки* | *доксограф* | *я* | *полиграф* | *оригинатор* | *автор этих строк* | *ткомедиограф* | *мы* | *песенник* | *комедиограф* | *авторикша*

Как можно заметить, в списке встречаются местоимения «я», «мы», занимающие 6 и 23 место по частоте встречаемости в русском языке. Естественно, появление таких частотных слов приводит к признанию их лучшими кандидатами в синонимы, а сам текст может превратиться в набор частотных слов. Для того, чтобы избежать этого, следует проводить фильтрацию кандидатов в синонимы, удаляя наиболее частотные слова (так называемые стоп-слова).

В списке синонимов встречаются не только отдельные слова, но и словосочетания. Это усложняет процедуры выбора синонимов – значения частотности появлений словосочетаний отсутствуют в [7], использование же данных других источников не позволит правильно соотносить частоты слов и словосочетаний в связи с отличиями в особенностях и коллекциях подсчета. Кроме того, значительно сложнее определить границы синонимимзации – отдельные слова или словосочетания, если словосочетания, то какой длины.

Заключение

Изложенный подход дает возможность проводить сравнение измененных текстов, в том числе с использованием метода шинглов, TF*IDF, сигнатур и т.д. [9]. При выполнении замены синонимов не нарушается структура текста и расположение входящих в него слов (сами слова могут при этом быть заменены). Эффект от описанного подхода будет достигаться только при обработке как оригинального, так и проверяемого текста.

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 19-07-00692.

Литература

1. Шабанова С.А. Сущность явлений синонимии и синонимизации // Новый университет. Серия: Актуальные проблемы гуманитарных и общественных наук, № 9, 2012, С. 48-50.
2. Чиркин Е.С. Системы автоматизированной проверки на неправомерные заимствования // Вестник Тамбовского университета. Серия: Гуманитарные науки. 2013. № 12 (128). С. 164-174.
3. Амирова О.Г. Роль контекста в процессе синонимизации слов // Проблемы лингвистики, методики обучения иностранным языкам и литературоведения в свете межкультурной коммуникации. Материалы II международной научно-практической конференции. 2009. С. 18-21.
4. Шумская А.О. Определение искусственных текстов на основе поиска часто употребляемых слов и устойчивых словосочетаний // Седьмая международная конференция по когнитивной науке. Тезисы докладов. 2016. С. 647-648.
5. Исхакова А.О. Метод и программное средство определения искусственно созданных текстов // автореферат диссертации на соискание ученой степени кандидата технических наук / Томский государственный университет систем управления и радиоэлектроники (ТУСУР) РАН. Томск, 2016
6. Исхакова А.О. Анализ текстовых признаков искусственных текстов, созданных на основе синонимизации // Научная сессия ТУСУР-2013. Материалы Всероссийской научно-технической конференции студентов, аспирантов и молодых ученых: в пяти частях. 2013. С. 224-226.
7. Зиберт А.О., Мирошниченко В.В. Об использовании словарей синонимов в алгоритме определения наличия заимствований в тексте // Universum: технические науки. 2014. № 12 (13). С. 3.
8. Ляшевская О.Н., Шаров С.А. Частотный словарь современного русского языка (на материалах Национального корпуса русского языка). - М.: Азбуковник, 2009. http://dict.ruslang.ru/freq.php
9. Sharapova E. One way to fuzzy duplicates detection // International Multidisciplinary Scientific GeoConference Surveying Geology and Mining Ecology Management, SGEM. 14. 2014. С. 273-278.