УДК 81.322.2

Проверка оригинальности синонимизированных текстов

Шарапова Е.В.

Синонимизация – замена слов в тексте синонимами (словами со схожим смыслом, но различным написанием). Основная цель синонимизации состоит в изменении текстового документа таким образом, чтобы повысить его уникальность, скрыв тем самым факт заимствования. В работе рассматриваются особенности проверки синонимизированных текстов и осуществляется поиск путей повышения качества выявления заимствований. Для обработки синонимизированных текстов предлагается использовать тяжелые синонимы (наиболее частотные, весомые синонимы). Проведенные исследования показали высокую эффективность подхода по сравнению с существующими системами проверки оригинальности. Одной из ключевых особенностей подхода является возможность использования различных алгоритмов информационного поиска для последующей обработки текста – «мешка слов», TF*IDF, N-грамм, шинглов и т.д. Это позволяет давать как статистическую оценку подобия проверяемых документов, так и проводить визуализацию найденных совпадений.

Ключевые слова: текст, оригинальности, синоним, синонимизация, оценка оригинальности, заимствование, антиплагиат.

Введение

Активное применение дистанционных форм обучения привело к росту числа работ, выполняемых студентами в электронной форме и сдаваемых на проверку. Для объективной оценки качества таких работ важным элементом является проверка их оригинальности, например, с помощью системы Антиплагиат. Далеко не всегда работы, сдаваемые студентами, имеют высокую оригинальность. Для ее повышения студенты пытаются использовать различные способы сокрытия – используют скрытый текст, замену символов, синонимизацию и т.д.

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 заключается в замене некоторых слов текста на синонимы. Синонимизация реализуется как вручную, так и в автоматическом режиме с помощью различных онлайн сервисов и программ (синонимайзеров) [3]. В настоящее время существует достаточно большое количество синонимайзеров – Raskruty.ru, Usyn.ru, Seogenerator.ru, Seobuilder.ru, Sinoni.men, Sinonimov.ru, Rustxt.ru, Textrobot.ru, Synonymizer.ru, Online-sinonim.ru, Progaonline.com/synonymizer/, Fromtlt.ru/sinonim и т.д. [4]. Они в значительной степени отличаются используемыми базами синонимов и алгоритмами работы (в первую очередь степенью переработки текста и его «читаемостью»).

Современные системы оценки оригинальности текстов (Антиплагиат, Text.Rucont.ru, Advego Plagiatus 3 и т.д.) видят в синонимизированных текстах значительную проблему – тексты ошибочно оцениваются системами как оригинальные, так как используемые алгоритмы чаще всего неспособны выявить подобные замены слов [5].

Цель работы – рассмотрение особенностей проверки синонимизированных текстов и поиск путей повышения качества выявления заимствований.

Оценка способности существующих систем обрабатывать синонимизированные тексты

Для оценки способности систем проверки текстов на оригинальность находить тексты, прошедшие синонимизацию, был сформулирован тестовый набор текстов с Wikipedia.org (то есть заведомо не уникальных, с оригинальностью 0%). Тексты прошли обработку в 12 системах синонимизации. Результаты

показали, что подавляющее число текстов были оценены системами проверки как уникальные (до 100% оригинальности). Худшие результаты продемонстрировали Fromtlt.ru/sinonim и Synonymizer.ru (оригинальность до 40%). Лучшие результаты у систем Raskruty.ru, Usyn.ru, Seogenerator.ru, Seobuilder.ru, Rustxt.ru и Progaonline.com/synonymizer/ (оригинальность более 90%).

Системы проверки оригинальности показывают различные результаты. При сильной переработке текстов (среднее значение оригинальности по системам более 95%: Raskruty.ru, Usyn.ru, Seogenerator.ru и Seobuilder.ru) меньшее значение оригинальности выдает система Pr-cy.ru/unique/. При средней глубине переработки текстов (среднее значение оригинальности по системам около 50%: Sinoni.men, Textrobot.ru) лучшие результаты показали Exactus.ru, Text.Rucont.ru, Text.ru, Content-watch.ru и Miralinks.ru. При незначительной переработке текстов (Fromtlt.ru/sinonim) неплохо справились системы Антиплагиат, Text.ru, Content-watch.ru, Advego.com/antiplagiat/, Advego Plagiatus 3, Text.Rucont.ru, Miralinks.ru и Exactus.ru.

Исследования показали, что синонимизация чаще всего позволяет повысить оригинальность текста и плохо обрабатывается существующими системами проверки. Тем не менее, имеет смысл обратить внимание на полученный после синонимизации текст. Анализ показывает, что текст после синонимизации становится трудно читаемым, содержит множество неправильно построенных фраз. Смысл некоторых предложений меняется до неузнаваемости.

Таблица 1. Сравнение эффективности работы систем проверки оригинальности

Системы проверки оригинальности текстов	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
Антиплагиат	100	100	100	100	92,9	100	100	66,8	96,7	47,4	100	14,8
Text.ru	100	100	100	100	28,1	42,6	60,8	35,8	100	19,7	100	14,6
Content-watch.ru	90,8	100	85,8	100	39,6	43,8	100	33,2	52,2	23,6	100	17,1
Pr-cy.ru/unique/	86	91	80	85	68	74	66	84	77	45	73	65
Advego.com/antiplagiat/	97	97	98	100	46	66	92	62	82	44	92	21
Advego Plagiatus 3	98	98	100	100	48	66	100	90	82	46	92	22
Etxt.ru	100	100	100	100	62	63	86	55	61	46	59	53
AntiPlagiarism.NET	98	98	98	98	75	79	95	74	88	53	97	48
Text.Rucont.ru	100	100	100	100	15	63	93	6	40,2	19	100	0
Be1.ru	100	100	100	100	100	100	100	100	100	85	100	94
Miralinks.ru	90	88	85	100	39	43	100	33	52	23	100	17
Exactus.ru	100	100	100	100	0	34,8	100	0	48,1	19,5	100	0
Среднее значение	96,7	97,7	95,6	98,6	51,1	64,6	91,1	53,3	73,3	39,3	92,8	30,5

Тяжелые синонимы

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 по словарю синонимов составляется список синонимов $w_i=\{s_{1,} s_{i2} \ldots s_{in}\}$ и по частотному словарю русского языка список их весов $f_i=\{f_{i1,} f_{i2} \ldots f_{in}\}$.

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 поиска 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 [5].

Таким образом, найденное наиболее представительное слово (с наивысшим весом) считается «тяжелым» синонимом исходного слова и используется для его замены.

Для работы с синонимами использовался словарь SynMaster, содержащий 1262190 позиций, и словарь Абрамова, содержащий 13709 позиций [6]. Надо заметить, что словарь Абрамова содержит слова, прошедшие лемматизацию (т.е. приведенные к нормальной форме), в то время как в словаре Synmaster содержатся различные словоформы, чем и объясняется его большой объем.

Было проведено 6 тестовых прогонов:

1. Без обработки синонимов,
2. Обработка с использованием тяжелых синонимов с удалением высокочастотных стоп-слов (с порогом частоты 500),
3. Обработка с использованием тяжелых синонимов без удаления стоп-слов,
4. Обработка с использованием самых тяжелых синонимов без удаления высокочастотных стоп-слов,
5. Обработка с использованием тяжелых синонимов по словарю Абрамова с удалением высокочастотных стоп-слов,
6. Обработка с использованием тяжелых синонимов по словарю Абрамова без удаления высокочастотных стоп-слов.

При первом прогоне не выполнялась обработка синонимов. Сравнение текстов проводилось в виде сравнения наборов входящих в них слов, приведенных к нормальной форме путем лемматизации [8].

При втором прогоне производилась замена слов текстов на их тяжелые синонимы [5]. При этом тяжелые синонимы определялись по словарю синонимов SynMaster итеративно с удалением высокочастотных стоп-слов. Порог частоты составлял 500. Надо заметить, что вариант с заменой на первый или случайный синоним из словаря был исключен при предварительном рассмотрении, как малоэффективный.

При третьем прогоне производилась замена слов текстов на их тяжелые синонимы, но без удаления высокочастотных стоп-слов из списка синонимов. Это привело к тому, что более 41% замен в словаре синонимов приходилось всего на 15 слов. Для сравнения, при удалении стоп-слов на первые 15 слов приходилось 26% замен. Надо заметить, что в частотном словаре Шарова [7] на первые 15 слов приходится около 27% всех упоминаний.

При четвертом прогоне производилась замена слов текстов на их самые тяжелые синонимы, найденные по списку всех синонимов в словаре. В отличие от предыдущих прогонов, список тяжелых синонимов формировался для всех слов словаря синонимов. Как результат, более 57% замен в словаре синонимов приходилось всего на 15 слов. Это неоправданно повышало близость текстов, существенно уменьшая при этом входящий в них набор слов.

Пятый прогон по параметрам был аналогичен второму прогону, но выполнялся по словарю синонимов Абрамова. Благодаря

качественному подходу к содержанию словаря, на первые 15 слов приходилось около 14% замен.

Шестой прогон по параметрам был аналогичен третьему прогону (замена слов текстов на их тяжелые синонимы без удаления высокочастотных стоп-слов из списка синонимов), но выполнялся по словарю синонимов Абрамова. На первые 15 слов приходилось не более 18% замен.

Анализ результатов прогонов показал, что лучшие результаты при проверке синонимизированных текстов показал прогон с использованием тяжелых синонимов. Немного уступают прогоны с использованием тяжелых синонимов без стоп-слов и самых тяжелых синонимов.

При использовании словаря Абрамова результаты получаются немного хуже, чем при использовании словаря SynMaster. Это можно объяснить меньшим размером словаря Абрамова, и, как следствие, меньшим охватом замен синонимов. Системы синонимизации используют в своей работе различные варианты замен слов, причем не всегда их синонимами. Иногда производится замена на слова, часто употребляемые в одном контексте, например, «датский» на «норвежский», «Отто» на «Ганс», «Китай» на «Иран» и т.д. Такие замены иногда позволяют правильно обработать более объемный словарь SynMaster, когда, например, и «Китай» и «Иран» могут быть заменены одним и тем-же тяжелым синонимом «страна».

Таблица 2. Результаты проверки синонимизированных текстов

Прогоны	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
Антиплагиат	100	100	100	100	92,9	100	100	66,8	96,7	47,4	100	14,8
Лучшая из систем	86	88	80	85	0	34,8	60,8	0	40,2	19,5	59	0
Среднее по системам	96,7	97,7	95,6	98,6	51,1	64,6	91,1	53,3	73,3	39,3	92,8	30,5
Без обработки синонимов	65,7	64,5	70,6	80,5	19,2	32,8	48,1	30,6	41,9	22,8	36,2	16,6
Тяжелые синонимы без стоп-слов	43,75	39,8	45,0	50,4	16,8	20,7	27,8	18,0	23,7	17,6	21,9	13,9
Тяжелые синонимы	28,5	26,5	27,3	34,9	16,0	12,9	22,3	7,2	16,3	11,3	23,2	7,7
Самые тяжелые синонимы	33,3	32,6	34,6	42,0	12,7	13,2	23,1	11,3	18,7	4,1	33,3	11,5
Тяжелые синонимы по словарю Абрамова без стоп-слов	53,7	50,2	58,0	64,2	16,5	24,6	38,6	24,4	31,3	17,8	31,8	14,6
Тяжелые синонимы по словарю Абрамова	50,9	46,7	51,7	62,2	15,6	24,6	38,7	23,6	31,0	14,7	30,2	12,1

Заключение

Полученные результаты проверки синонимизированных текстов нашей системой в несколько раз превосходят результаты проверки существующими системами. Так, система Антиплагиат в 9 случаях из 12 определила оригинальность синонимизированных текстов более 90%. При использовании тяжелых синонимов в 8 случаях оригинальность составила менее 25%. При этом наивысшее значение оригинальности составило менее 35%.

Применение тяжелых синонимов для обработки синонимизированных текстов позволяет приблизиться к решению проблемы завышенной оценки оригинальности текстов. Одной из ключевых особенностей подхода является возможность использования различных алгоритмов информационного поиска для последующей обработки текста – «мешка слов», TF*IDF, N-грамм, шинглов и т.д. Это позволяет давать как статистическую оценку подобия документов, так и проводить визуализацию найденных совпадений.

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. *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.

3. *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.

4. Текстовые заимствования и борьба с ними: монография / *Е. В. Шарапова*; Владим. гос. ун-т им. А. Г. и Н. Г. Столетовых. – Владимир: Изд-во ВлГУ, 2020. – 160 с.

5. *Середа С.Н., Шарапова Е.В.* Борьба с синонимизацией в текстовых потоках данных / Прикладные вопросы формирования и обработки сигналов в радиолокации, связи и акустике [Электронный ресурс]: Всероссийские открытые Армандовские чтения «Муром 2020» / Сб. тез. докладов XI научно-практического семинара. – Муром: Изд.-полиграфический центр МИ ВлГУ, 2020. – С.7-9.

6. *Абрамов Н.* Словарь русских синонимов и сходных по смыслу выражений. – М.: Русские словари, 1999.

7. *Ляшевская О.Н., Шаров С.А.* Частотный словарь современного русского языка (на материалах Национального корпуса русского языка). – М.: Азбуковник, 2009.

8. *Усков И.В.* Лемматизация русских текстов компьютером / Автоматизация, мехатроника, информационные технологии. Материалы III Международной научно-технической интернет-конференции молодых ученых. 2013. С. 182-185.

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

Поступила 15 сентября 2020 г.

Synonymization is the replacement of words in a text with synonyms (words with a similar meaning, but different spellings). The main purpose of synonymization is to change a text document in such a way as to increase its uniqueness, thereby hiding the fact of borrowing. The paper discusses the features of checking synonymized texts and searches for ways to improve the quality of detecting borrowings. For the processing of synonymized texts, it is proposed to use heavy synonyms (the most frequent, weighty synonyms). The studies carried out have shown the high efficiency of the approach in comparison with the existing systems for checking originality. One of the key features of the approach is the ability to use various information retrieval algorithms for subsequent text processing - a “bag of words”, TF*IDF, N-grams, shingles, etc. This allow to give both a statistical assessment of the similarity of documents and visualize the found matches.

Key words: text, originality, synonym, synonymization, assessment of originality, borrowing, anti-plagiarism.

Шарапова Екатерина Викторовна – старший преподаватель кафедры техносферной безопасности Муромского института (филиала) Федерального государственного бюджетного образовательного учреждения высшего образования "Владимирский государственный университет имени Александра Григорьевича и Николая Григорьевича Столетовых".

E-mail: sharapovamivlgu@gmail.com.

602264, г. Муром, ул. Орловская, д. 23.