Шарапова Е.В.
Муромский институт (филиал) федерального государственного образовательного учреждения высшего образования «Владимирский государственный университет имени Александра Григорьевича и Николая Григорьевича Столетовых»
602264, г. Муром, Владимирская обл., ул. Орловская, 23
E-mail: sharapovamivlgu@gmail.com

Обнаружение синонимизированных заимствованных текстов с помощью «тяжелых» синонимов*

Одним из способов сокрытия заимствований текстов в настоящее время является синнонимизация, то есть замена слов в тексте синонимами (словами со схожим смыслом, но различным написанием) [1]. При этом текст меняется таким образом, что системы проверки считают оценивают его как уникальный и не определяют наличие заимствований [2]. Синонимизация представляет собой лексическую замену некоторых слов текста их синонимы. Такая замено проводится в ручном (перефразирование) и автоматическом режиме с использованием так называемых синонимайзеров (онлайн сервисов или программ). 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») [3].

В настоящее время проблему обнаружения и обработки синонимизации предлагается решать путем расширения запросов сравнения синонимами (например, с использованием WordNet) [4, 5, 6]. Другим решением является сравнение словарей (наборов слов) проверяемых текстов, расширенных с помощью синонимов и очищенных от часто употребляемых слов [7, 8]. Подобные решения имеют один существенный недостаток – они работают с текстом как набором слов и не позволяют учитывать их взаимное расположение. По этой причине возникают проблемы с визуализацией найденных совпадений и корректного подсчета заимствований.

Проблему можно решить с помощью изложенного ниже подхода с использованием «тяжелых» синонимов. Для этих целей будет использоваться три словаря:

1. Словарь термов (слов) коллекции документов W. Для этих целей используется словарь системы Автор.NET (более 160 тысяч слов) [9].
2. Частотный словарь русского языка F [10].
3. Словарь синонимов S (база синонимов SynMaster).

Рассмотрим процедуру определения «тяжелых» синонимов:

1. Для каждого слова w_i в словаре W поставим в соответствие его вес f_i, подсчитанный на основе глобальной частоты его встречаемости в русскоязычных текстах. При отсутствии значения в таблице глобальных частот вес принимается равным 0.
2. Для каждого слова w_i по словарю S составляется список синонимов $w_i=\{s_1, s_{i2} \ldots s_{in}\}$.
3. Для каждого синонима подсчитываются их веса $f_i=\{ f_{i1}, f_{i2} \ldots f_{in}\}$.
4. Определяется наиболее представительный («тяжелый») синоним в соответствии с максимальным значением веса, т.е. $\max(f_{i1}, f_{i2} \ldots f_{in}) \rightarrow s_i$.
5. 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 (переходим на шаг 7).
6. Синоним-кандидат s_i считается как исходное слово w'_i и действие итерационно повторяется для него (переходим к шагу 2).
7. Слово w_i заменяется найденным синонимом s_i. При весе слова больше весов всех кандидатов на синонимы, его замена синонимами не производится, а слово считается наиболее представительным.

Как можно заметить, процедура поиска синонимов заключается в выявлении синонимов с максимальным весом, затем для найденных синонимов также ищется синоним с максимальным

весом и так далее, до тех пор, пока не будет найден синоним, вес у которого больше всех остальных кандидатов другими словами самый «тяжелый» синоним.

Основная идея подхода заключается в том, что какой бы из синонимов не был выбран в момент синонимизации (и из какой базы синонимов), он вероятнее всего попадет в рассматриваемый список синонимов. Так как мы не знаем, какое именно слово было заменено на какой синоним, мы можем заменить все слова ни их самые вероятные («тяжелые») синонимы. В этом случает и исходный и проверяемый (прошедший синонимазацию для сокрытия заимствований) текст будут изменены примерно одинаково на одни и те же «тяжелые» синонимы. То есть не нарушается структура текста и расположение входящих в него слов (сами слова могут при этом быть заменены). Поэтому сравнение их содержания даст вполне приемлемые результаты.

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

Синоним	Вес
я	15631.11
мы	4740.29
писатель	213.5
автор	160.56
композитор	17.93
литератор	16.04
создатель	13.04
творец	12.36
виновник	9.3
сочинитель	4.71
составитель	2.2
полиграф	2.2
либреттист	0
компилятор	0
доксограф	0
оригинатор	0
ткомедиограф	0
песенник	0
комедиограф	0
авторикша	0

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 нашем случае был выбран порог веса в 500, что обеспечивает удаление 190 наиболее частотных слов русского языка. Таким образом, «тяжелым» синонимом слова «автор» будет слово «писатель».

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. Шарапова Е.В. 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.

4. Chong, M., and Specia, L. 2011. Lexical Generalisation for Word-level Matching in Plagiarism Detection. In RANLP, pp. 704-709.

5. Nawab, R. M. A., Stevenson, M., and Clough, P. 2016. An ir-based approach utilising query expansion for plagiarism detection in medline. IEEE/ACM transactions on computational biology and bioinformatics.

6. Nawab, R. M. A., Stevenson, M., and Clough, P. 2012, April. Retrieving candidate plagiarised documents using query expansion. In European Conference on Information Retrieval. Springer Berlin Heidelberg, pp. 207-218

7. Atadjanov J. Document Plagiarism Detection On The Internet With Accounting Synonym Forms Of Words // International journal of scientific and technology research, vol. 8, issue 12, 2019, 1517-1519.

8. 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

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.

10. 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

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