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

Оценка качества работы сервисов проверки оригинальности текстов

При разработке системы проверки оригинальности текстов Автор.НЕТ [1, 2] возникла задача оценки качества ее работы и сравнения результатов работы с существующими системами.

Для оценки качества определения полных дубликатов системами проверки оригинальности текстов был сформирован тестовый набор. В него вошли русскоязычные тексты с Wikipedia.org. Тексты никак не модифицировались. Таким образом, оригинальность текстов из тестового набора составляла 0 %. Результаты проверки различными системами приведены в табл.1.

Как можно заметить, большинство систем оценило оригинальность от 0 до 1%. Разница вызвана, видимо, особенностями реализации различных систем и способами подсчета оригинальности (по символам, по словам). Система Be1.ru показала оригинальность менее 10%. Вероятно, такая выдача результата заложена в алгоритм системы. Система Etxt.ru показала оригинальность 19%, необоснованно признав некоторые куски текста за уникальные.

Таблица 1. Поиск полных дубликатов

Система	Оригинальность, %	
	Wikipedia.org	Новости
Антиплагиат	0,19	74,32
Text.ru	0,45	2,36
Content-watch.ru	1	33,2
Pr-cy.ru/unique/	2	10
Advego.com/antiplagiat/	0	4
Advego Plagiatus 3	0	4
Etxt.ru	19	52
AntiPlagiarism.NET	0	3
Text.Rucont.ru	0	100
Be1.ru	10	22
Miralinks.ru	1	33
Exactus.ru	0	100
Автор.НЕТ	0	2,9

Таким образом, подавляющее большинство систем уверенно справились с задачей поиска неоригинальных текстов. Учитывая, что все рассмотренные системы проверки по собственной базе работ индексируют – результаты получились хорошие.

Далее, мы взяли свежие сообщения с новостных сайтов. В данном случае, новостные сообщения не попали в базы работ систем Антиплагиат, Text.Rucont.ru и Exactus.ru. Как следствие, Антиплагиат определил оригинальность документов как 74,32%, а Text.Rucont.ru и Exactus.ru – как 100%. Достаточно слабые результаты показала система Etxt.ru, определив оригинальность новостных сообщений как 52%. Лучшие результаты показали AntiPlagiarism.NET, Advego Plagiatus 3, Advego.com/antiplagiat/ и Text.ru менее 4%.

Таким образом, использование только внутренней базы работ без поиска в сети Интернет делает системы Антиплагиат, Text.Rucont.ru и Exactus.ru малопригодными для оценки оригинальности текстов любой тематики и «свежести» написания. В более ранних исследованиях была выявлена неспособность системы Антиплагиат находить ряд дубликатов текстов, опубликованных в местной прессе и региональных сайтах [3, 4].

Система Автор.НЕТ правильно определила оригинальность текстов, взятых из Wikipedia.org как 0, а новостей 2,9, немного уступив лидеру Text.ru.

Для оценки поиска нечетких дубликатов был составлен текст, содержащий 50% оригинального текста и 50% взятого с Wikipedia.org. Число слов и символов в каждом из указанных частей было одинаковым. Оригинальные и неоригинальные предложения были перемешаны между собой. Полученный таким образом текст был проанализирован различными системами проверки оригинальности (табл. 2). Результаты показали достаточно высокую точность всех систем. Наиболее близкие к 50% результаты показали системы Антиплагиат (50,25%) и Advego Plagiatus 3 (51%). Система Автор.НЕТ определила оригинальность в 50,5, что близко к лидеру – системе Антиплагиат.

Надо заметить, что разница в результатах оценки оригинальности может быть с связана с особенностями подсчета процента совпадений разными системами. Так, при малом проценте совпадений системы могут округлять его до 0, а при большом – до 100%. Кроме того, сама методика подсчета совпадений в системах отличается – по словам, по символам, с учетом или без учета знаков препинания и стоп-слов. По этой причине, разницу в 1-2% вполне можно списать на особенности реализации той или иной системы.

Таблица 2. Поиск нечетких дубликатов

Система	Оригинальность, %
Антиплагиат	50,25
Text.ru	48,16
Content-watch.ru	51,9
Pr-cy.ru/unique/	48
Advego.com/antiplagiat/	52
Advego Plagiatus 3	51
Etxt.ru	53
AntiPlagiarism.NET	52
Text.Rucont.ru	48
Be1.ru	54
Miralinks.ru	52
Exactus.ru	48,02
Автор.НЕТ	50,5

Таким образом, точность работы системы Автор.НЕТ в большинстве случаев превосходит точность существующих систем, делая ее достойным конкурентом.

Литература

1. Шарапова Е.В., Шарапов Р.В. Система проверки текстов на заимствования из других источников // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: XIII Всероссийская научная конференция «RCDL'2011». Воронеж, 19-22 октября 2011 г.: труды конференции – Воронеж: Издательско-полиграфический центр

2. Шарапова Е.В., Шарапов Р.В. Обнаружение нечетких дубликатов текстов в больших массивах информации // Проблемы управления и моделирования в сложных системах: Труды XXI Международной конференции (3-6 сентября 2019 г. Самара, Россия). – Самара: ООО “Офорт”, 2019. Том 2. С. 335-339.

3. Шарапова Е.В. Исследование возможностей системы "Антиплагиат" для обнаружения заимствований // Перспективы науки и образования, 2013, № 3. – С. 215-219.

4. Шарапова Е.В., Шарапов Р.В. Исследование плагиата в работах студентов // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной международной конференции «Диалог 2012» (Бекасово, 30 мая – 3 июня 2012 г). Вып. 11 (18). Том 1 – М: Изд-во РГГУ, 2012. – С. 578-586.