УДК 519.25; 004.8

А.О. Шумская

Выбор параметров для идентификации искусственно созданных текстов

Проведено исследование искусственных текстов в сравнении с естественными, на основе которых они были созданы. Определены численные текстовые характеристики, которые в большей степени изменяются при использовании методов искусственной генерации текстов и могут быть использованы для идентификации (распознавания) искусственно созданных текстов.
Ключевые слова: текст, авторство, искусственные тексты, идентификация, характеристики текста.

Проблема идентификации искусственных текстов. Идентификация искусственно сгенерированных текстов имеет важное прикладное значение для области знаний, связанной с информационной безопасностью и построением систем защиты информации. Исследования этого направления особенно актуальны в связи с увеличением объема текстовых массивов, появлением новых способов их распространения в компьютерных сетях, увеличением случаев анонимности и плагиата.

В связи с этим требуется проведение исследований и поиск новых решений, способных дать ответ на вопрос, был ли текст написан человеком (естественный текст) или он создан искусственно. Искусственными (искусственно созданными) текстами называются текстовые произведения, сгенерированные специальными программами-генераторами.

Подходы к идентификации происхождения текста. Под идентификацией понимается процесс установления происхождения объекта по совокупности общих и частных признаков, образующих так называемый авторский стиль [1, 2, 6]. Авторский стиль в методах атрибуции текста представляется в виде *авторского инварианта*, который определяется совокупностью численных значений текстовых характеристик. На основе инварианта можно выявить причастность текста к какому-либо автору либо группе авторов.

Для идентификации искусственно созданных текстов вне зависимости от автора оригинального произведения либо вообще при отсутствии такового, предлагается создание авторского инварианта текстов, созданных с помощью методов автоматической генерации.

Предлагается следующий подход к решению задач определения искусственных текстовых форм:

– экспериментальное исследование характеристик искусственных текстов и определение параметра или набора параметров, присущих конкретным методам автоматического создания текста;

– отбор наиболее информативных параметров и выбор состава авторского инварианта искусственных текстов;

– определение эффективности идентификации искусственных текстов с помощью предложенного авторского инварианта;

– исследование нескольких частей одного текста и сравнение их характеристик с целью выявления изменений авторского стиля.

Для подобного рода исследований большое значение имеет объем текстовой формы: в коротком сообщении трудно выделить логические части и корректно рассчитать показатели текста, что может повлечь ошибки в идентификации.

Идентифицирующие характеристики искусственных текстов. Для определения авторства широко используются статистические методы исследования авторских инвариантов [3]. В работах, посвященных атрибуции текстовых форм [4], выделяются следующие характеристики, составляющие авторский инвариант:

– массовость, под которой понимается свойство параметра слабо контролироваться автором на сознательном уровне;

– устойчивость – сохранение значения параметра в некотором диапазоне для одного автора;

– различающая способность, т.е. способность текстовой характеристики принимать существенно отличающиеся значения для разных авторов. Существенное различие понимается как превышение диапазона разброса значений для текстов одного автора.

Для проведения расчетов и выявления идентифицирующих признаков искусственных текстов были выбраны следующие текстовые характеристики: количество предложений в тексте, количество служебных слов, средняя длина слова, упоминание определенных слов, количество коротких слов, количество длинных слов.

Указанные характеристики были рассчитаны как для сгенерированных текстов, так и для работ, автор которых заведомо известен. Генерация искусственных текстов проводилась с помощью свободно распространяемых программных продуктов SyMonym и Article Clone Easy, которые основаны на синонимизации исходного текста.

Расчеты характеристик искусственных текстов. С помощью указанных выше программ были созданы два искусственных варианта текста известного автора. В обеих программах использовался заявленный способ синонимизации, однако в каждой из них был собственный словарь синонимов. Используемый словарь играет важную роль при синонимизации, от него зависит качество и, что самое главное, при создании искусственных текстовых форм – уникальность сгенерированного текста.

С помощью специального онлайн-сервиса [5] были оценены уровни схожести оригинального произведения и искусственно созданных вариантов. Выбранный метод реализует наиболее распространенный алгоритм для вычисления уровня схожести – алгоритм Шинглов. Данный алгоритм основан на приведении текстов к канонической форме (удаление служебных слов, приведение слов к единой падежной форме и т.д.) и сравнении слов и/или словосочетаний текста методом контрольных сумм. Процент уникальности в данном случае рассчитывался как разность: 100% – «процент схожести текстов».

Процент уникальности варианта SyMonym с оригинальной версией текста в среднем составил 22,03%. Для текста, сгенерированного программой Article Clone Easy, процент уникальности составил в среднем 68,71%. Очевидно, что показатели уникальности заметно отличаются.

Были исследованы три варианта одного и того же произведения: оригинальный – публицистический текст по психологии; вариант, сгенерированный программой SyMonym; вариант, сгенерированный программой Article Clone Easy. Результаты расчетов, нормированные к максимальному значению, приведены в таблице и на рис. 1.

Результаты расчета текстовых характеристик

Параметры	Тексты		
	Оригинал	Версия SyMonim	Версия Article Clone Easy
Предложения	1	0,9743	0,8926
Служебные слова	1	0,9834	0,7310
Ср. длина слова	0,9410	0,9474	1
Упоминание определенных слов	1	0,9804	0,5686
Короткие слова	1	0,9870	0,7489
Длинные слова	0,9831	0,9837	1

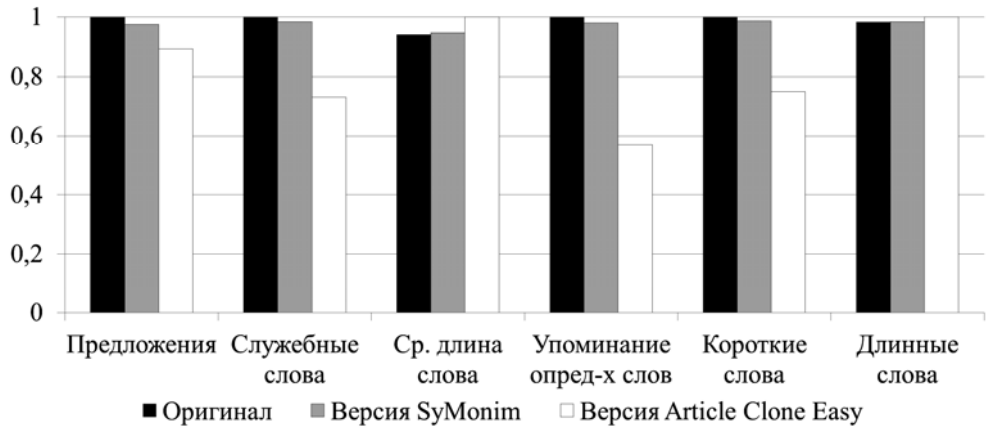

Рис. 1. Текстовые характеристики исследуемых выборок

Приведенные данные показывают, что после преобразования текста в нем увеличились количество длинных слов и средняя длина слова. Количество предложений и упоминаний искомых слов незначительно уменьшилось за счет того, что в процессе преобразования текста были добавлены

новые слова. При анализе текстов длиной 1000 знаков эти показатели уменьшились на доли процента. Сам метод генерации текста не вносит изменений в его грамматическую структуру.

Разность параметров текстов тем больше, чем больше указанная выше степень их уникальности. Это свидетельствует о том, что синонимизация действительно влияет на использованные текстовые параметры и проявляется в изменении их значений тем более, чем большее количество замен было произведено в тексте. Видны и параметры, значения которых незначительно изменяются – например количество длинных слов.

Значимость полученных изменений текстовых характеристик может быть проверена с помощью критериев согласия, таких как критерий Пирсона, критерий Стьюдента, расстояние Махаланобиса и др.

Заключение. Важнейшим элементом процесса атрибуции текста является определение вектора идентифицирующих признаков, образующего авторский инвариант. Признаки должны обладать свойствами массовости, устойчивости и различающей способности.

Состав вектора признаков в каждом конкретном случае зависит от уровня исследования текста, определяется перечисленными свойствами признаков, а также может зависеть от специфики поставленной задачи. Проведенные расчеты позволили выделить характеристики, показательные для синонимизации, и характеристики, слабо изменяющиеся при этом методе генерации.

В качестве показательных характеристик синонимизации можно выделить количество служебных слов, упоминание определенных слов, количество коротких слов. Слабо изменяемыми характеристиками являются количество длинных слов и средняя длина слова. Таким образом, был получен материал для апробации других искусственных текстов, а также для расчета новых параметров, прямо или косвенно связанных с характеристиками, наиболее изменяемыми в приведенных расчетах.

Направлениями дальнейших исследований являются проверка статистических гипотез различия естественных и искусственных текстов, а также изучение особенностей других алгоритмов генерации текста, в том числе основанных на Марковских цепях, использовании словаря и др.

Литература

1. Романов А.С. Разработка и исследование математических моделей, методик и программных средств информационных процессов при идентификации автора текста / А.С. Романов, А.А. Шелупанов, Р.В. Мещеряков. – Томск : В-Спектр, 2011. – 188 с.

2. Романов А.С. Методика идентификации автора текста на основе аппарата опорных векторов // Доклады ТУСУРа. – 2009. – № 1 (19), ч. 2. – С. 36–42.

3. Родионова Е.С. Методы атрибуции художественных текстов // Структурная и прикладная лингвистика : межвуз. сб. / под ред. А.С. Герда. – СПб. : Изд-во СПб. гос. ун-та, 2008. – Вып. 7. – С. 118–127.

4. Романов А.С. Состояние проблемы распознавания и идентификации автора текста // Информационная безопасность [Электронный ресурс] – Режим доступа: http://inf-bez.ru/?p=813

5. SEObuilding.ru поисковая оптимизация [Электронный ресурс] // Определение похожих текстов. Сравнение текстов на схожесть – [Б.м.], 2013. – Режим доступа: http://www.seobuilding.ru/similar-text-checker.php. (дата обращения: 12.03.2013).

6. Технология прямого поиска при решении задач прикладной математики / В.А. Архипов, С.С. Бондарчук, И.Г. Боровской, А.А. Шелупанов // Вычислительные технологии. – 1995. – Т. 4, № 10. – С. 19.

Шумская Анастасия Олеговна
Инженер каф. КИБЭВС ТУСУРа
Тел.: 8-952-804-00-69
Эл. почта: shumskaya.ao@gmail.com

Shumskaya A.O.

Choice of parameters for identification of artificial texts

The article presents the research of automatically generated texts in comparison with naturally established. Numerical characteristics of the text are identified. The results can be used to develop a system of identification of automatically generated texts.

Keywords: text, authorship, automatically generated, identification of the text characteristics.