УДК 004.072.7 **DOI** 10.15622/sp.49.6

А.О. Шумская

МЕТОД ОПРЕДЕЛЕНИЯ ИСКУССТВЕННЫХ ТЕКСТОВ НА ОСНОВЕ РАСЧЕТА МЕРЫ ПРИНАДЛЕЖНОСТИ К ИНВАРИАНТАМ

Шумская А.О. **Метод определения искусственных текстов на основе расчета меры принадлежности к инвариантам.**

Аннотация. Работа посвящена вопросу идентификации текстов, сгенерированных автоматически (искусственно) с помощью программных алгоритмов. Актуальность данной задачи обусловлена ростом распространения таких текстов в Интернете. Создаваемые «копии» веб-страниц используются для привлечения читателей к интернет-ресурсам, а также для распространения большого количества уникальных экземпляров страниц с контентом определенной направленности.

В статье описаны особенности определения происхождения текста на примере работы с текстами, созданными с помощью синонимизации как наиболее распространенного метода генерации искусственных текстов, представляющих собой веб-контент. Предложен инвариант искусственно созданных текстов, представляющий собой набор значений текстовых характеристик, который позволяет классифицировать тексты по способу их создания. Предложен метод определения искусственно созданных текстов на основе расчета меры принадлежности входного текста к инвариантам, позволяющий принять решение о происхождении текста. В статье также приведены значения, полученные в ходе проведения серии экспериментов по определению искусственно созданных текстов.

Ключевые слова: автоматически сгенерированные тексты, искусственные тексты, массовое порождение текстов, текстовые характеристики, атрибуция текста.

1. Введение. Задачи, связанные с атрибуцией текстов, имеют широкий прикладной характер: они направлены на разрешение литературоведческих вопросов, выявление фактов нарушения авторского права, проведение социальных исследований, а также криминалистических и иных расследований.

Способы распространения текстовой информации определяют направление исследований в этой области. За последнее десятилетие компьютерные сети стали важнейшим инструментом для обмена данными. Все большее место в общественных отношениях и взаимодействиях стали занимать интернет-ресурсы. Информация, распространяемая в сети, зачастую воспринимается человеком как современный аналог энциклопедии или справочников, а также телевидения и газет, которым люди привыкли верить. Массово порожденные с помощью программных алгоритмов уникальные тексты могут быть использованы для введения пользователей в заблуждение, распространения заведомо ложной или подстрекающей информации и в других неправомерных целях [1].

Методы массового автоматического порождения текстов, которые применяются для SEO-оптимизаций и в других сферах, являются также инструментом для создания множества «уникальных» копий определенного текста [2]. Уникальность достигается благодаря использованию специальных алгоритмов, связанных с изменением текстообразующих элементов. Одним из наиболее часто используемых методов создания искусственных текстов является синонимизация, при которой происходит изменение текста путем замены отдельных лексем на схожие по смыслу (синонимы). Для реализации данного метода чаще всего используются программы-синонимайзеры, а также соответствующие словари синонимов, которые в совокупности позволяют автоматизировать процесс создания текстов.

Существующие решения в области определения текстов, сгенерированных автоматически, касаются в первую очередь выявления поискового спама, к которому относят: дублирование веб-страниц, создание дорвеев, манипуляции с текстом сайта с использованием алгоритмов массового порождения текстов. Паразитное воздействие веб-спама направлено в первую очередь на поисковые системы, с чем связаны и механизмы, используемые для его создания и распространения.

Например, в работах А. С. Павлова и А. М. Райгородского с соавторами [3–5] предложен алгоритм определения текстового спама на основе оценки разнообразия тематик документа. Метод основан на оценке характеристик текста, отличающих поисковый спам, к которым относятся обилие ключевых слов, наличие скрытого текста, мелкого нечитаемого шрифта, злоупотребление тегами заголовков и т.д. Использование данных характеристик не позволяет использовать метод для определения следов автоматического генератора на текстах, распространяемых как информационный контент, неблагоприятное воздействие которого направлено не на поисковые системы, а непосредственно на пользователей веб-ресурсов.

Также исследования веб-спама опубликованы зарубежными авторами: C. Castillo, D. Donato и др. [6]. В основе предложенного ими метода — анализ веб-контента на наличие ссылок на определенные страницы и их распределения между страницами. В своей работе они приводят схемы взаимодействия ссылок для паразитных спам-страниц и аналогичные — для нормальных страниц. Данный метод обнаружения автоматически сгенерированного текста также имеет в своей основе особенности веб-спама и неприменим для иных текстов.

В работе А. А. Зайцева с соавторами [7] рассмотрен метод оценки качества текста на основе реферирования. Метод предполагает, что тексты, созданные с применением синонимизаторов или средств авто-

матического перевода, не обладают достаточным тематическим единством. Это выражается в том, что используемые в тексте лексемы не соответствуют одной общей теме, имеются словосочетания, не соответствующие тематике. По этой причине при повышении компрессии «сжатия» реферата у искусственно созданных текстов с большей скоростью уменьшается его размерность. Авторами выведен ряд эвристических правил, которые позволяют исключить тексты низкого качества, тем самым доказывая, что тематические признаки текста позволяют выявить связные, информативные тексты.

Определение текстов, представляющих собой машинный перевод с одного естественного языка на другой — еще одна задача идентификации происхождения текста. Для выявления таких текстов существует ряд специальных алгоритмов, например, BLEU, METEOR, Perplexity, оценивающих качество машинного перевода через соответствие характеристик текста на лексическом уровне некоторым эталонным показателям естественного языка. Данные алгоритмы опробованы в ряде работ для решения различных прикладных задач. Например, в работе R. Aharoni с соавторами [8] предлагается расширить алгоритм BLEU рассмотрением текстов также на уровне предложения. Авторами показано, что для некоторых языков данный подход дает более точные показатели метрики BLEU.

В результате анализа рассмотренных методов и подходов были сделаны следующие выводы:

1) Существующие методы по определению неестественных текстов позволяют решать задачи выявления поискового спама, а также текстов, являющихся машинным переводом текста с одного естественного языка на другой.

2) Из-за наложенных ограничений и используемых особенностей рассмотренных текстов существующие методы не могут быть использованы в задаче определения текстов, созданных с помощью синонимизаторов или других средств автоматической генерации.

3) Для достижения большей точности в определении связности текста необходимо рассматривать все лингвистические уровни текста, в том числе синтаксический.

4) Качество текста напрямую связано с его тематическими свойствами, то есть связанные с ними характеристики текста должны быть включены в инварианты.

5) При формировании набора текстовых характеристик, используемых для идентификации неестественных текстов, следует отталкиваться от глобальных свойств текстов.

Таким образом, необходимо разработать метод определения текстов, сгенерированных автоматически и представляющих собой информационный веб-контент. При решении данной задачи необходимо учесть достижения исследований по оценке качества текстовых произведений и определения поискового спама, а также в области выделения характерных черт естественных текстов [9], анализа интернет-сообщений [10–13] и автоматической атрибуции текстов [14–18], которые косвенно относятся к решению задачи определения массово порожденных текстов.

В статье приводится описание и результат исследования по определению происхождения текстов на основе исследования статистических значений характеристик текста. В работе были исследованы тексты, являющиеся контентом социальных сетей и веб-ресурсов для коммуникации различных сообществ людей.

2. Метод определения происхождения текста. Исходными данными в задаче классификации текстов по способу их создания являются:

– бесконечное множество T текстов, которые могут быть исследованы;

– множество X исследуемых текстовых характеристик признаков $X = \{x_1, x_2, \ldots, x_m\}$, $|X| = m$;

– множество инвариантов классов текстов, разделенных по происхождению — наборов усредненных значений тестовых характеристик $x_j \in X$ для n классов текстов: $A = \{a_1, a_2, \ldots, a_n\}$, $a_i = (a_{i1}, a_{i2}, \ldots, a_{im})$, где a_{ij} — усредненное информативное значение j-ой текстовой характеристики i-го инварианта, $i = 1..n$, $j = 1..m$.

Таким образом, задача определения происхождения сводится к задаче отнесения входного текста к одному из классов текстов с известным происхождением, то есть к классификации. Для формализации данной задачи вводятся дополнительные обозначения:

– a' — набор рассчитанных значений текстовых характеристик некоторого входного текста t, происхождение которого требуется определить, $t \in T$;

– a_j'— рассчитанное значение j-ой текстовой характеристики входного текста, $j = 1..m$;

– $V(a', A)$ — мера оценки принадлежности входного текста к классу текстов с известным происхождением;

– $D(a', a_i)$ — мера расстояния между входным текстом и i-ым классом текстов известного авторства, представляемая как мера расстояния между векторами: a' и a_i;

– l — пороговое значение расстояния между вектором значений текстовых характеристик входного текста a' и вектора-инварианта i-го

класса текста с известным происхождением a_i такое, что значение максимальное значение меры $D(a', a_i)$ не должно превышать l при i=1..n.

Входной текст, представленный в виде вектора a', считается принадлежащим тому классу текстов, расстояние до которого является минимальным, но не превышающим заданного порогового значения l. Тогда определение происхождения текста есть задача, которую формально можно описать следующим образом:

$$V(a', A) = \min[D(a', a_i)];\ i = 1...n;$$
$$a' = (a'_1, a'_2, ..., a'_m);$$
$$a_i = (a_{i1}, a_{i2}, ..., a_{im}).$$

Принимается решение о том, что входной текст отнесен к i-му классу, если выполняется выражение:

$$V(a', A) \equiv D(a', a_i);$$
$$V(a', A) \le l.$$

Пороговое значение l ограничено сверху величиной, равной половине меры расстояния между двумя наиболее приближенными векторами-инвариантами a_x, $a_y \in A$ и может быть скорректировано в меньшую сторону на основе экспериментальных расчетов:

$$l \le \frac{D(a_x; a_y)}{2}.$$

Для принятого решения об отнесении входного текста к некоторому классу на заключительном шаге метода рассчитывается точность данного заключения R_A:

$$R_A = 1 - \frac{V(a', A)}{D(a_x, a_y)},$$

где a_x, $a_y \in A$ — два вектора-инварианта, наиболее приближенных (согласно мере расстояния) к a'.

Значение R_A лежит в интервале [0;1] и оно тем выше, чем меньше расстояние между вектором значения характеристик входного текста с вектором-инвариантом, с которым он был соотнесен.

Входной текст не может быть отнесен ни к одному из классов текстов с известным происхождением, если выполняется:

$$\forall a_i \in A \quad D(a', a_i) > l;\ i = 1..n.$$

Для расчета меры расстояния между векторами a' и a_i может быть использована любая метрика, позволяющая количественно оценить расстояние между двумя точками в m-мерном пространстве. Среди возможных метрик — общеизвестные Евклидова метрика и метрика Махаланобиса. Нужно отметить, что выбор метрики должен отвечать требованиям поставленной задачи.

Так как для решения описанной задачи необходим единый масштаб получаемых значений меры $D(a', a_i)$, для ее расчета предлагается использовать метрику (расстояние) Махаланобиса [19]. Данная метрика обобщает понятие расстояния Евклида, учитывает корреляции между переменными и инвариантно к масштабу. Она широко используется в кластерном анализе и методах классификации.

Используя метрику Махаланобиса, меру расстояния между векторами a' и a_i можно представить следующим выражением:

$$D(a', a_i) = \sqrt{(a' - a_i)^T \cdot S^{-1} \cdot (a' - a_i)},$$

где S — объединенная ковариационная матрица.

При этом значения меры будут строго принадлежать интервалу: $D(a', a_i) \in [0; 1]$. Также значение величины l должно лежать в этом же интервале: $l \in [0; 1]$.

3. Инварианты искусственных и естественных текстов. Важнейшей частью исследования является формирование инвариантов разных по происхождению текстов [20, 21]. Это обусловлено тем, что на их основе в дальнейшем принимается решение о способе создания текста. В процессе формирования инвариантов важно учитывать принципы, принятые для атрибуции в целом: характеристики должны быть массовыми, а их значения — устойчивыми и обладать различительной способностью [22]. Формирование инварианта для решения задачи определения происхождения текстов — процесс, состоящий из нескольких этапов:

1) определение лингвистических признаков, по которым различимы тексты разных классов;

2) определение групп свойств текста, на которые влияют выделенные лингвистические признаки;

3) формирование перечня характеристик текста, которые позволяют измерить проявление свойства текста для каждой из групп;

4) проведение расчетов значений характеристик текста для тестовых наборов текстов известного происхождения;

5) проверка различительной способности характеристик;

6) проверка взаимозависимости характеристик;

7) удаление характеристик, не обладающих различительной способностью, а также коррелирующих с другими в наборе;

8) формирование инварианта как набора значений выделенных характеристик текста.

Отличием естественных текстов от искусственных является их связность в рамках межфразовых единств, а также цельность, то есть наличие глобальной связи компонентов текста на содержательном уровне. Таким образом, связность и цельность являются непременными лингвистическими признаками текста, которые проявляются в целесообразно построенном человеком тексте и отличают его от массово порожденных экземпляров [23, 24]. Эти факты также были подтверждены автором субъективно при рассмотрении сгенерированных текстов: при замене слов синонимами или перемене их мест нарушается связность словосочетаний, предложений, абзацев, теряется смысл, тематические свойства текста.

Во внимание было принято, что учеными-лингвистами выделено [23] пять основных свойств текста, обеспечивающих его связность и цельность:

– символьные (связанные с наличием букв, буквосочетаний, цифр, символов пунктуации, математических символов);

– лексические (связанные с наличием слов и словосочетаний, словарное разнообразие);

– синтаксические (связанные с конструкциями предложений в тексте, синтаксическими особенностями текста);

– семантические (связанные с оценкой мер семантического сходства и связности текстообразующих средств);

– тематические (связанные с соответствием используемых текстообразующих средств тематике текста).

Для выделенных свойств сформированы наборы характеристик текста. Нужно отметить, что в системах атрибуции чаще всего используются символьные и лексические свойства текстов [25, 26]. Это, в первую очередь, обусловлено вычислительной простотой расчетов, связанных с определением значений текстовых характеристик. В свою очередь, синтаксические и семантические анализаторы, а также тематические рубрикаторы предъявляют дополнительные требования к разметке текста и зависят от качества используемых справочников. Алгоритмы анализа таких свойств характеризуются высокой вычислительной сложностью, так как менее формализованы и зависят от многообразия форм языка, а также могут изменяться с течением времени.

Для формирования инвариантов классов текстов были использованы 3210 текстов и созданные на их основе автоматические сгене-

рированные экземпляры. Генерация искусственных текстов производилась с использованием метода синонимизации. Данные тексты представляли собой публицистические статьи информационного характера длиной от 1000 до 5700 символов. В качестве генератора был использован синонимизатор со словарем, содержащим 700 тысяч синонимов. В общей сложности при оценке численных значений характеристик текстов двух классов были использованы 3210 естественных текстов общим объемом 10,7 млн. символов и 3210 искусственных текстов (12,3 млн. символов). Указанные объемы считаются достаточными для обучения в соответствии с опытом формирования инвариантов при анализе текстов различных авторов, приведенных в работах [27].

Так как набор исследуемых характеристики был сформирован на основе предположения, что естественный текст отличается от искусственного цельностью и связностью, ожидаемым оказался результат, что 81% характеристик показали значительное изменение: от 25 до 60%, тем самым показав свою различительную способность. Синтаксические свойства показали незначительные или неопределенные изменения, несмотря на то, что синтаксис текста является основой его связности. Такие результаты могут быть связаны с особенностями механизма генерации текста. Характеристики, не проявившие различительную способность, были исключены из перечня. В ходе корреляционного анализа были также исключены характеристики, обладающие высокой корреляционной зависимостью. Окончательный набор состоял из следующих характеристик текста:

- среднее количество знаков пунктуации;
- частоты 100 популярных биграмм букв;
- частота служебных слов;
- количество уникальных слов;
- среднее число слов в предложении;
- количество грамматических ошибок;
- количество предложений;
- количество сложноподчиненных предложений;
- количество вопросительных предложений;
- частота 100 популярных слов;
- частота 100 популярных 2-грамм слов;
- количество слов в семантическом ядре;
- наличие единства тематики в разных частях текста.

В результате были сформированы два инварианта: для естественных текстов, то есть созданных человеком, и для искусственных, созданных с помощью синонимизации. Инварианты, как было сказано ранее, представляют собой векторы значений. Размерность таких век-

торов соответствует количеству исследуемых характеристик текста. Ниже приведены полученные векторы численных значений характеристик текста *(*a_1 — инвариант класса естественных текстов; a_2 — инвариант класса искусственных текстов, сгенерированных с помощью синонимизации):

$$a_1 = (\ 31{,}742;\ 201{,}269;\ 34{,}691;\ 64{,}804;\ 9{,}113;\ 0{,}01;\ 109{,}812;\ 68{,}655;\ 1{,}414;\ 49{,}001;\ 9{,}1;\ 66{,}025;\ 1{,}7);$$

$$a_2 = (29{,}035;\ 112{,}562;\ 25{,}702;\ 101{,}659;\ 9{,}987;\ 6{,}215;\ 100{,}2;\ 62{,}082;\ 1{,}358;\ 32{,}882;\ 3{,}554;\ 95{,}645;\ 0{,}6).$$

4. Эксперименты с расчетом меры принадлежности. Для оценки эффективности предложенного метода и сформированных инвариантов была проведена серия экспериментов по расчету меры оценки принадлежности некоторых входных текстов известного происхождения к классам текстов с известным происхождением.

Эксперименты проводились для 1000 интернет-статей, собранных автором в популярных социальных сетях. Статьи отбирались по следующим тематическим направленностям: общество, политика, финансы, власть, армия, силовые структуры, наука и техника и смежные с ними. Общий объем исходной выборки естественных текстов составил 1,965 млн символов.

Естественные тексты подверглись действию автоматического генератора на основе синонимизации для того, чтобы создать искусственные экземпляры. Для дополнительной оценки влияния объема словаря синонимов и, как следствие, уникальности искусственного текста на результат определения происхождения текста были использованы 2 словаря: с 700 тысячами синонимов и с 130 тысячами синонимов. Среднее значение уникальности искусственных текстов, созданных с помощью словаря с меньшим объемом, рассчитанное с помощью алгоритма шинглов [28], составило 36,5%, тогда как для словаря с большим объемом это значение составило 69,1%.

Таким образом, в эксперименте использовано 3 выборки по 1000 текстов, объединенных одним из способов создания:

– естественные тексты, написанные человеком или несколькими людьми;

– искусственные тексты, созданные с помощью синонимизации на основе словаря из 700 тысяч синонимов;

– искусственные тексты, созданные с помощью синонимизации на основе словаря из 130 тысяч синонимов.

Примеры искусственных текстов, созданных с помощью синонимизатора, приведены в таблице 1.

Таблица 1. Примеры искусственных текстов

Оригинал (естественный текст)	Искусственные тексты, созданные с помощью синонимизатора	
	Словарь содержит 180 тысяч синонимов	Словарь содержит 700 тысяч синонимов
Полагаю, через десять лет экономика России будет такой же, как сейчас. Никаких радикальных изменений в её структуре мы, к сожалению, не увидим.	Полагаю, сквозь десять лет экономика России довольно такой же, словно теперь. Никаких радикальных изменений в её структуре мы, к сокрушению, не увидим.	Считаю, посредством 10 года макроэкономика Российской федерации станет такого рода ведь, равно как в настоящее время. Практически никаких конструктивных перемен в ее текстуре я, к огорчению, никак не узнаем.
Это плохо, потому что мировая конъюнктура цен на нефть, газ и другие сырьевые ресурсы меняется, и не в нашу пользу. Это приведет к тому, что Россия к 2026 году резко опустится в ранге стран с развитой экономикой.	Это низко, потому который сделка конъюнктура валют для нефть, газ и другие сырьевые ресурсы меняется, и не в нашу выгоду. Это приведет к тому, который Россия к 2026 году явно опустится в ранге территорий с образованный экономикой.	Это слабо, вследствие того то что всемирная ситуация стоимости в черное золото, голубое топливо и прочие сырьевые средства изменяется, и никак не в нашу с тобой выгоду. Данное повергнет к этому, то что Российская федерация к 2026 г. сильно снизится в ранге государств с сформированной экономикой.
Некоторые улучшения возможны разве что в сельском хозяйстве, IT, отдельных узких нишах машиностроения. Но на общем фоне деградации это не создаст нового качества экономики.	Некоторые исправления допустимы неужели который в сельском хозяйстве, IT, частичных тесных нишах машиностроения. Однако для всеобщем фоне деградации это не создаст свежего качества экономики.	Некоторые усовершенствования вероятны неужели то что в аграрном хозяйстве, IT, единичных ограниченных нишах машиностроения. Однако в совокупном фоне деградации данное никак не сформирует новейшего особенности экономики.
За счёт развития сельского хозяйства имеют шансы укрепиться Краснодарский край, Ставрополье, однако это также не скажется на общем состоянии экономики страны.	За счёт развития сельского хозяйства заключают шансы овладевать Краснодарский кромка, Ставрополье, впрочем это также не скажется для всеобщем положении экономики территории.	За счёт формирования аграрного хозяйства обладают возможности закрепиться Краснодарский область, Ставрополье, но данное кроме того никак не отразится в совокупном пребывании экономики государства.

Показатели уникальности представленных текстов, рассчитанные с помощью алгоритма шинглов, указаны в таблице 2. Приведенные показатели позволяют сделать вывод, что объем словаря синонимов влияет на уникальность создаваемых текстов: чем полнее словарь, тем меньше сходство текста с оригиналом и, следовательно, выше его уникальность.

Таблица 2. Показатели уникальности искусственных текстов в зависимости от полноты используемых словарей синонимов

Размер словаря синонимов		180 тысяч синонимов	700 тысяч синонимов
Сходство искусственного текста с оригиналом	для шингла из 1 слова	49 %	17 %
	для шингла из 2 слов	24 %	3 %
	для шингла из 3 слов	13 %	0 %
Уникальность созданного искусственного текста		51 %	83 %
Вероятное поведение поисковой системы при индексировании страницы с полученным искусственным текстом		Страница с данным текстом, вероятно, не будет проиндексирована, возможна блокировка ресурса	Страница с данным текстом, вероятно, будет успешно проиндексирована

Вероятное поведение поисковых систем относительно индексации тех или иных текстов, приведенное в таблице 2, было спрогнозировано на основании данных, которые открыто распространяет компания Яндекс [29], однако точные алгоритмы работы поисковых машин, в частности критерии блокировки веб-страниц, скрыты.

5. Результаты расчетов. В статье приведены результаты экспериментальных расчетов для 30 первых текстов из каждой выборки, которые иллюстрируют процедуру принятия решения на основе метода в зависимости от значений рассчитанных расстояний между вектором значений входного текста и векторами-инвариантами известных классов текстов.

Для публикации результатов экспериментальных расчетов использовались следующие обозначения, соответствующие описанным выше:

a_1 — инвариант класса естественных текстов;

a_2 — инвариант класса искусственных текстов, сгенерированных с помощью синонимизации;

A — множество инвариантов. $A = \{ a_1, a_2 \}$.

Для каждого входного текста, происхождение которого необходимо было определить, был рассчитан вектор a' значений характеристик.

В соответствии с принятыми ранее обозначениями и приведенными формулами мера расстояния между входным текстом и i-ым

классом текстов известного авторства $D(a', a_i)$ рассчитывалась как мера расстояния Махаланобиса между векторами:a' и a_i.

Результаты расчетов меры расстояния до каждого класса и результирующая мера принадлежности для каждого случая приведены в таблице 3.

Таблица 3. Значения мер близости, меры принадлежности входных текстов

a' (п/п)	$D(a', a_1)$	$D(a', a_2)$	$V(a', A) \equiv D(a', a_i)$	Заключение о происхождении текста
На входе — естественные тексты (созданные человеком или несколькими людьми)				
1	0,121	0,311	$a_i = a_1$	Естественный
2	0,058	0,480	$a_i = a_1$	Естественный
3	0,214	0,501	$a_i = a_1$	Естественный
4	0,125	0,398	$a_i = a_1$	Естественный
5	0,174	0,413	$a_i = a_1$	Естественный
6	0,270	0,407	$a_i = a_1$	Естественный
7	0,116	0,457	$a_i = a_1$	Естественный
8	0,041	0,579	$a_i = a_1$	Естественный
9	0,149	0,325	$a_i = a_1$	Естественный
10	0,240	0,306	$a_i = a_1$	Естественный
На входе — искусственные тексты, созданные с помощью синонимизации (словарь, содержащий 180 тысяч синонимов)				
11	0,140	0,307	$a_i = a_1$	Естественный
12	0,240	0,201	$a_i = a_2$	Искусственный
13	0,298	0,231	$a_i = a_2$	Искусственный
14	0,056	0,121	$a_i = a_1$	Естественный
15	0,220	0,078	$a_i = a_2$	Искусственный
16	0,306	0,101	$a_i = a_2$	Искусственный
17	0,380	0,356	$a_i = a_2$	Искусственный
18	0,160	0,114	$a_i = a_2$	Искусственный
19	0,405	0,330	$a_i = a_2$	Искусственный
20	0,280	0,197	$a_i = a_2$	Искусственный
На входе — искусственные тексты, созданные с помощью синонимизации (словарь, содержащий 700 тысяч синонимов)				
21	0,350	0,079	$a_i = a_2$	Искусственный
22	0,307	0,104	$a_i = a_2$	Искусственный
23	0,345	0,056	$a_i = a_2$	Искусственный
24	0,221	0,116	$a_i = a_2$	Искусственный
25	0,432	0,032	$a_i = a_2$	Искусственный
26	0,401	0,074	$a_i = a_2$	Искусственный
27	0,390	0,140	$a_i = a_2$	Искусственный
28	0,374	0,099	$a_i = a_2$	Искусственный
29	0,340	0,067	$a_i = a_2$	Искусственный
30	0,259	0,012	$a_i = a_2$	Искусственный

В таблице 4 приведены показатели ошибок 1-го и 2-го рода на основе полученных результатов проведенных вычислений для полных выборок. Под ошибками 1-го рода понимаются случаи, когда естественный текст был принят за искусственный (ложноположительное событие, или «ложная тревога» для пользователя системы). Ошибки 2-го рода указывают на случаи, когда искусственный текст не был распознан системой и был принят за естественный экземпляр (ложноотрицательное событие, или «пропуск события»).

Таблица 4. Показатели ошибок 1 и 2 рода

Показатель	Ошибки 1-го рода	Ошибки 2-го рода
Определение естественного текста	3,2%	1,8%
Определение текста, созданного с помощью синонимизации (словарь из 180 тысяч синонимов)	11,6%	6,1%
Определение текста, созданного с помощью синонимизации (словарь из 700 тысяч синонимов)	4,2%	2,8%

Таким образом, на основе произведенных экспериментальных вычислений с участием выборок текстов известного происхождения были определены показатели эффективности. Ошибки 1 рода составили 4,2%, ошибки 2 рода — 2,8%. Следует отметить, что метод показал большую эффективность при определении искусственных текстов, которые обладают уникальностью выше 60%, то есть созданных с помощью более полного словаря синонимов. Большое количество ошибок в определении происхождения текстов, созданных с использованием 130 тысяч синонимов, объясняется тем, что алгоритм слабо изменил исходный текст и, как следствие, искусственный текст не был сгенерирован полноценно и не обладал достаточной уникальностью.

5. Заключение. Исследование текстов, сгенерированных автоматически — новый виток области знаний, связанных с текстовой атрибуцией. Идентификация таких текстов имеет ряд особенностей по сравнению, например, с задачей определения авторства, которые связаны, в первую очередь, с тем, что исходный материал искусственного текста может быть написан любым автором, группой авторов или уже быть продуктом программного алгоритма.

Автором предложен метод определения происхождения текста на основе статистических расчетов средних значений текстовых характеристик, использующий меру близости двух векторов в m-мерном пространстве как основание к отнесению спорного текста к одному из из-

вестных классов. Экспериментальные расчеты, проведенные с использованием текстов, сгенерированных методом синонимизации, показали эффективность в принятии решения с помощью предложенного метода.

Литература

1. Управление ООН по наркотикам и преступности. Использование интернета в террористических целях. С. 3–6. URL: https://www.unodc.org/ documents/terrorism/ Publications/Use_of_Internet_for_Terrorist_Purposes/Use_of_ the_ internet_for_terrorist_ purposes_Russian.pdf (дата обращения: 26.05.2016).
2. SEO-копирайтинг: как приручить поисковик. URL: http://onedesign.pro/upload/ books/11_Kak_priruchit.pdf (дата обращения: 01.06.2016).
3. *Павлов А.С., Добров Б.В.* Методы обнаружения поискового спама, порожденного с помощью цепей Маркова // Тр. XI Всероссийский научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции». 2009. Т. 1. С. 311–317.
4. *Павлов А.С., Добров Б.В.* Метод обнаружения массово порожденных неестественных текстов на основе анализа тематической структуры // Вычислительные методы и программирование. 2011. Т. 12. Вып. 3. С. 58–72.
5. *Гречников Е.А., Гусев Г.Г., Кустарев А.А., Райгородский А.М.* Поиск неестественных текстов // Труды XI всероссийской конференции «Цифровые библиотеки: продвинутые методы и технологии, цифровые коллекции» – RCDL'2009. Петрозаводск. 2009. С. 306–308.
6. *Castillo C., Donato D., Becchetti L., Boldi P., Leonardi S., Santini M., Vigna S.* A reference collection for web spam // ACM Sigir Forum 2006. 2006. vol. 40. Issue 2. pp. 11–24.
7. *Зайцева А.А., Кулешов С.В., Михайлов С.Н.* Метод оценки качества текстов в задачах аналитического мониторинга информационных ресурсов // Труды СПИИРАН. 2014. Вып. 37. С. 144–155.
8. *Aharoni R., Koppel M., Goldberg Y.* Automatic Detection of Machine Translated Text and Translation Quality // Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. 2014. vol. 2. pp. 289–295.
9. *Анисимов А.В., Лиман К.С., МарченкоА.А.* Методы вычисления мер семантической близости слов естественного языка // Искусственный интеллект. 2010. №3. С. 170–175.
10. *Brocardo M.L., Traore I., Saad S. Woungang I.* Authorship Verification for Short Messages using Stylometry // Journal of Computer and System Sciences. 2015. vol. 91. Issue 8. pp. 1429–1440.
11. *Zheng R., Li J., Chen H., Huang Z.* A Framework for Authorship Identification of Online Messages: Writing-Style Features and Classification Techniques // Journal of the American society for information science and technology. 2006. vol. 57 Issue 3. pp. 378–393.
12. *Ragel R.G., Herath P., Senanayake U.* Authorship detection of SMS messages using unigrams // Industrial and Information Systems (ICIIS): 8th IEEE International Conference. 2013. pp. 387–392.
13. *Brocardo M.L., Traore I., Saad S. Woungang I.* Verifying Online User Identity using Stylometric Analysis for Short Messages // Journal of networks. 2014. vol. 9. no. 12. pp. 3347–3355.
14. *Романов А.С.* Методика идентификации автора текста на основе аппарата опорных векторов // Доклады ТУСУР. 2009. №1(19). Часть 2. С. 36–42.

15. *Романов А.С., Резанова З.И., Мещеряков Р.В.* Методика проверки однородности текста и выявления плагиата на основе метода опорных векторов и фильтра быстрой корреляции // Доклады ТУСУР. 2014. № 2(32). С. 264–269.
16. *Романов А.С., Шелупанов А.А., Мещеряков Р.В.* Разработка и исследование математических моделей, методик и программных средств информационных процессов при идентификации автора текста: Монография. Томск: В-Спектр. 2011. 188 с.
17. Лингвоанализатор. URL: www.rusf.ru/books/analysis (дата обращения: 26.05.2016).
18. *Corney M., Anderson A., Mohay G., de Vel O.* Identifying the Authors of Suspect Email. URL: http://eprints.qut.edu.au/8021/1/CompSecurityPaper.pdf (дата обращения 26.05.2016).
19. *Шумская А.О.* Оценка эффективности метрик расстояния Евклида и расстояния Махаланобиса в задачах идентификации происхождения текста // Доклады ТУСУРа. 2013. № 3(29). С. 141–145.
20. *Шумская А.О.* Идентифицирующие признаки текстовых сообщений при установлении автора // Ползуновский вестник. 2013. № 2. С. 265–266.
21. *Шумская А.О.* Выбор параметров для идентификации искусственно созданных текстов // Доклады ТУСУРа. 2013. № 2(28). С. 126–128.
22. *Фоменко В.П., Фоменко Т.Г.* Авторский инвариант русских литературных текстов. Приложение: кто был автором «Тихого Дона»? Дополнение 3 к книге Фоменко А.Т. «Методы статистического анализа исторических текстов (приложения к хронологии)» в 2-х томах. М.: Крафт+Леан, 1999. 832+908 с.
23. *Валгина Н.С.* Теория текста. М.: Логос. 2003. 191 с.
24. *Николина Н.А.* Филологический анализ текста: учеб. Пособие // М.: Издательский центр "Академия". 2003. 256 с.
25. *Напреенко Г.В.* Идентификация текста по его авторской принадлежности на лексическом уровне (формально-количественная модель) // Вестник Томского государственного университета. 2014. № 379. С. 17–23.
26. *Красса С.И.* Методика и инструментарий атрибуции текста в автороведческой экспертизе // Альманах современной науки и образования. 2013. № 10(77). С. 106–108.
27. *Романов А.С., Шелупанов А.А., Бондарчук С.С.* Обобщенная методика идентификации автора неизвестного текста // Доклады ТУСУР. 2010. № 1(21). Часть 1. С. 108–112.
28. *Зеленков Ю.Г., Сегалович И.В.* Сравнительный анализ методов определения нечетких дубликатов для Web-документов. URL: http://rcdl2007.pereslavl.ru/ papers/ paper_65_v1.pdf (дата обращения: 01.06.2016).
29. Некачественные сайты: Чем отличается качественный сайт от некачественного с точки зрения Яндекса? URL: https://yandex.ru/support/webmaster/yandex-indexing/webmaster-advice.xml (дата обращения: 30.09.2016).

Шумская Анастасия Олеговна — аспирант кафедры комплексной информационной безопасности электронно-вычислительных систем, Томский государственный университет систем управления и радиоэлектроники (ТУСУР). Область научных интересов: исследование искусственных текстов, автоматическая идентификация искусственных текстов, нейронные сети, искусственный интеллект, статистические методы обработки текстов, прикладная лингвистика. Число научных публикаций — 15. shumskaya.ao@gmail.com; пр. Ленина, 40, Томск, 634050; р.т.: +79138142824.

DOI 10.15622/sp.49.6

A.O. SHUMSKAYA

METHOD OF THE ARTIFICIAL TEXTS IDENTIFICATION BASEN ON THE CALCULATION OF THE BELONGING MEASURE TO THE INVARIANTS

Shumskaya A.O. **Method of the Artificial Texts Identification based on the Calculation of the Belonging Measure to the Invariants.**

Abstract. The work is devoted to the identification of texts generated automatically (artificially) with the use of software algorithms. This is an important and topical issue, because such texts are being widely spread on the Internet. Created «copies» of the web pages are used to attract readers to online resources as well as to disseminate a large number of unique copies of pages with content specific orientation.

This article describes the features of determining the origin of the text by the example of working on texts generated by synonymization as the most common method of generating artificial web content. The author provides an invariant of artificial texts as a set of the values of text characteristics, which allows classification of texts according to the process of their creation. The article proposes a method of the artificial texts identification based on the calculation of the belonging measure to the invariants, which allows making a decision about the origin of the text. The article also presents values obtained from the experiments on identifying artificial texts.

Keywords: automatically generated texts, artificial texts, massively generated texts, text features, text attribution.

Shumskaya Anastasia Olegovna — Ph.D. student of complex security of electronic-computing systems department, Tomsk State University of Control Systems and Radioelectronics (TUSUR). Research interests: automatically generated texts identification, computational linguistics, neural networks, artificial intelligence, word processing statistical methods. The number of publications — 15. shumskaya.ao@gmail.com; 40, Leninavenue, Tomsk, 634050, Russia; office phone: +79138142824.

References

1. Upravlenie OON po narkotikam i prestupnosti. Ispol'zovanie interneta v terroristicheskih celjah [United Nations Office on Drugs and Crime. The use of the Internet for terrorist purposes]. pp. 3–6. Available at: https://www.unodc.org/documents/terrorism/ Publications/Use_of_Internet_for_Terrorist_Purposes/Use_of_the_internet_for_terrorist_ purposes_Russian.pdf (accessed 26.05.2013). (In Russ.).
2. SEO-kopirajting: kak priruchit' poiskovik [SEO-copyrighting. How you can taim the search engine]. Available at: http://onedesign.pro/upload/books/11_Kak_priruchit.pdf (accessed 01.06.2016). (In Russ.).
3. Pavlov A.S., Dobrov B.V. [Methods for detection of Web Spam Created With Markov Chains]. *Tr. XI Vserossijskij nauchnoj konferencii «Jelektronnye biblioteki: perspektivnye metody i tehnologii, jelektronnye kollekcii* [Proceedings of the XI conference "Digital libraries: advanced methods and technologies, digital collections"]. 2009. Issue 1. pp. 311–317. (In Russ.).
4. Pavlov A.S., Dobrov B.V. [Method of detection mass generated unnatural texts by analyzing the thematic structure]. *Vychislitel'nye metody i programmirovanie – Numerical methods and programming*. 2011. Issue 12. vol. 3. pp. 58–72. (In Russ.).

5. Grechnikov E.A., Gusev G.G., Kustarev A.A., Rajgorodskij A.M. [Unnatural texts search]. *Trudy XI vserossijskoj konferencii «Cifrovye biblioteki: prodvinutye metody i tehnologii, cifrovye kollekcii» – RCDL'2009* [Proceedings of the XI Conference "Digital Libraries: Advanced Methods and Technologies"]. Petrozavodsk. 2009. pp. 306–308. (In Russ.).
6. Castillo C., Donato D., Becchetti L., Boldi P., Leonardi S., Santini M., Vigna S. A reference collection for web spam. ACM Sigir Forum 2006. vol. 40. Issue 2. 2006. pp. 11–24.
7. Zaytseva A.A., Kuleshov S.V., Mikhailov S.N. [The Method for the Text Quality Estimation in the Task of Analytical Monitoring of Information]. *Trudy SPIIRAN – SPIIRAS Proceedings*. 2014. vol. 37. pp. 144–155. (In Russ.).
8. Aharoni R., Koppel M., Goldberg Y. Automatic Detection of Machine Translated Text and Translation Quality. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. 2014. vol. 2. pp. 289–295.
9. Anisimov A.V., Liman K.S., Marchenko A.A. [Methods for calculating measures of natural language words semantic similarity]. *Iskusstvenniy intellect – Artificial Intelligence*. 2010. vol. 3. pp. 170–175. (In Russ.).
10. Brocardo M.L., Traore I., Saad S. Woungang I. Authorship Verification for Short Messages using Stylometry. *Journal of Computer and System Sciences*. 2015. vol. 91. Issue 8. pp. 1429–1440.
11. Zheng R. Li J., Chen H., Huang Z. A Framework for Authorship Identification of Online Messages: Writing-Style Features and Classification Techniques. *Journal of the American society for information science and technology*. 2006. vol. 57. Issue 3. pp. 378–393.
12. Ragel R.G., Herath P., Senanayake U. Authorship detection of SMS messages using unigrams. Industrial and Information Systems (ICIIS): 8th IEEE International Conference. 2013. pp. 387–392.
13. Brocardo M.L., Traore I., Saad S. Woungang I. Verifying Online User Identity using Stylometric Analysis for Short Messages. *Journal of networks*. 2014. vol. 9 no. 12. pp. 3347–3355.
14. Romanov A.S. [Methods of identifying the author of the text , based on support vector machine]. *Doklady TUSUR – Proceedings of TUSUR University*. 2009. vol. 1(19). Part 2. pp. 36–42. (In Russ.).
15. Romanov A.S., Rezanova Z.I., Meshcheryakov R.V. [Methodology for testing homogeneity of the text and plagiarism detection method based on support vector machines and fast correlation filter]. *Doklady TUSUR – Proceedings of TUSUR University*. 2014. vol. 2(32). pp. 264–269. (In Russ.).
16. Romanov A.S., Shelupanov A.A., Meshcheryakov R.V. *Rasrabotka I issledovanie matematicheskikh modeley, metodik i programmnikh sredstv informationnikh processov pri identifikacii avtora teksta: Monographiya*. [Development and research of mathematical models, methods and software for information processes in the text author identification: Monograph]. Tomsk: V-Spektr. 2011. 188 p. (In Russ.).
17. Lingvoanalizator [Lingvoanalisator]. Available at: www.rusf.ru/books/analysis (accessed 26.05.2016). (In Russ.).
18. Corney M., Anderson A., Mohay G., de Vel O. Identifying the Authors of Suspect Email. URL: http://eprints.qut.edu.au/8021/1/CompSecurityPaper.pdf (accessed 26.05.2016).
19. Shumskaya A.O. [The effectiveness of Euclidean distance and the Mahalanobis distance in the problems of identification of the text origin]. *Doklady TUSUR – Proceedings of TUSUR University*. 2013. vol. 3(29). pp. 141–145. (In Russ.).

20. Shumskaya A.O. [Identification features of the text messages in the establishment of the author]. *Polzunovskiy vestnik – Polzunovsky vestnik*. 2013. vol. 2. pp. 265–266. (In Russ.).
21. Shumskaya A.O. [Choice of parameters for identification of artificial texts]. *Doklady TUSUR – Proceedings of TUSUR University*. 2013. vol. 2(28). pp. 126–128. (In Russ.).
22. Fomenko V.P., Fomenko T.G. *Avtorskij invariant russkih literaturnyh tekstov. Prilozhenie: kto byl avtorom «Tihogo Dona»? Dopolnenie 3 k knige Fomenko A.T. «Metody statisticheskogo analiza istoricheskih tekstov (prilozhenija k hronologii)» v 2-h tomah* [Russian literary texts author invariant. Appendix : Who was the author of «And Quiet Flows the Don»? Supplement 3 to the book of Fomenko A.T. «The methods of statistical analysis of historical texts (annex to the chronology)» in 2 volumes]. Moscow.: Kraft+Lean. 832+908 p. (In Russ.).
23. Valgina N.S. *Teoriya teksta* [The theory of text]. M.: Logos. 2003. 191 p. (In Russ.).
24. Nikolina N.A. *Philologicheskiy analis teksta: uchebnoe posobie* [Philological analysis of text: tutorial]. M.: Isdatelskiy centr “Akademiya". 2003. 256 p. (In Russ.).
25. Napreenko G.V. [Authorship identification of the text on the lexical level (formal-quantitative model)]. *Vestnik Tomskogo gosudarstvennogo universiteta – Tomsk State University Journal*. 2014. vol. 379. pp. 17–23. (In Russ.).
26. Krassa S.I. [Methodology and instruments of text attribution in authorship expertise]. *Almanah sovremennoj nauki i obrazovanija – Almanac of Modern Science and Education*. 2013. vol. 10(77). pp. 106–108. (In Russ.).
27. Romanov A.S., Shelupanov A.A., Bondarchuk S.S. [Generalized authorship identification technique]. *Doklady TUSUR – Proceedings of TUSUR University*. 2014. vol. 1(21). Part 1. pp. 108–112. (In Russ.).
28. Zelenkov Yu.G., Segalovich I.V. Comparative analysis of methods for near-duplicate detection for Web-documents. Available at: http://rcdl2007.pereslavl.ru/ papers/ paper_65_v1.pdf (accessed 01.06.2016). (In Russ.).
29. Nekachestvennye sajty: Chem otlichaetsja kachestvennyj sajt ot nekachestvennogo s tochki zrenija Yandexa? [Low-quality websites: What is the difference between qualitative and low-quality sites from the point of view of Yandex?]. Available at: https://yandex.ru/support/webmaster/yandex-indexing/webmaster-advice.xml (accessed 30.09.2016). (In Russ.).