УДК 004.838.2

Краткий обзор методов выявления искусственных текстов

Д.Д. Береснева

[1]Московский физико-технический институт (государственный университет)

Искусственными (искусственно созданными) текстами называются текстовые произведения, сгенерированные специальными программами-генераторами. 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.

Целью данного обзора является отбор методов определения искусственного или модифицированного машиной текста для их сравнения и последующей реализации согласно обозначенной исследователем проблеме. Эта задача актуальна для выявления фальшивых работ среди научных публикаций, а так же для проблемы нахождения поискового спама.

В дальнейших исследованиях предполагается проверка статистических гипотез различия естественных и автоматически сгенерированных х текстов, а также изучение особенностей различных алгоритмов генерации текста, в том числе основанных на Марковских цепях, использовании словаря и др.

Для обзора были отобраны 96 статей из Интернета, удовлетворяющем тематике распознавания модифицированного машиной текста. В качестве источника данных использовались такие ресурсы, как Google Shcolar, Elibrary.ru, dl.acm.org. В конечную выборку попало 14 отобранных статей. В данной работе приводятся некоторые из методов, рассмотренных в статьях.

Метод корреляций

Для обнаружения принадлежности текста к автоматически сгенерированным или написанным человеком используется исследование корреляций соседних слов в тексте авторы hgtlkfuf.n следующий метод [1].

Берутся 2000 самых часто используемых слов русского языка.

Aij - матрица 2000x2000, в которой на пересечении i-й строки и j-го столбца стоит частота

встречаемости в языке пары слов с номерами i и j (вычисляется по фиксированной базе текстов ruscorpa.

Функция $Cor(i, j) = \frac{A_{ij}}{A_i} + \frac{A_{ij}}{A_j}$ измеряет степень «сочетаемости» слов с номерами i и j,

где A_i и A_j – суммы по строкам и столбам.

Метод лингвистических особенностей

Другим способом [2] выявления машинного перевода выступает метод, использующий лингвистические и статистические характеристики текста. Метод не зависит от исходного языка. Данными послужили 350000 выровненных Англо-Испанских пар предложений, взятых из инструкций и мануалов и онлайн-справок.,200K предложений на английском для обучения лингвистических моделей (эталон) и 100K выровненных пар предложений. Используются различные индексы и характеристики для анализа: индексы - скаляры, вектора - нормализируются по длине предложения.. Далее строятся деревья решений, основанные на характеристиках, полученных из оригинальных и искусственных текстов.

Для избегания переобучения узлы в деревьях решений не разветвляются, если они насчитываются в менее чем 50 случаях.

Метод, использующий SVM-алгоритм [4]

Для обучения используется Canadian Hansard корпус англо-французских текстов. Данные разбивались на 4 кластера: английские настоящие, английские, переведенные машиной, французские настоящие и французские, переведенные машиной. Из текстов извлекались частоты юниграмм, приведенные к длине каждого документа. Документы рассматривались полностью;отдельные абзацы не рассматривались. Для каждого документа отношение TTR нестэммированных юниграмм рассчитывалось как средняя длина юниграмм; Эти дополнительные функции были добавлены, чтобы смоделировать лексические упрощения характеристик искусственного текста. Юниграммы, представления числа и символы были удалены, оставлены только слова и лексемы. Документы, содержащие менее 20 лексем, были удалены. Далее используется SVM-алгоритм для классификации.

Метод, использующий Hidden Style Similarity

Выявление автоматически сгенерированных (шаблоном или скриптом) текстов в Интернете данным [3] способом происходит с помощью «меры скрытой схожести» текстов на основе вне текстовых особенностей в исходном HTML коде. Также строится алгоритм кластеризации для сортировки сгенерированных текстов в зависимости от метода генерации. В качестве данных берётся корпус из 5 миллионов HTML страниц, отобранных поисковым роботом по заданным критериям. При определенном пороге метод дает 100% точность исследования.

Определение спама: 1) новые отдельные случаи уже обнаруженного спама (контролируемо) 2) выявление набора подозрительных страниц из большом корпусе без категории (сложно контролируемо).

Метод группирует веб-страницы на основе шаблона и стиля письма. Это позволяет обнаружить конкретный жанр, как, например, страницы форумов или каталоги веб-серверов, и исключить их из результатов поиска. Учитывая набор уже известных "нежелательных" страниц, другие найденные страницы, подходящие под шаблон нежелательных, будут удаляться.

Метод расчета энтропии [14]

Для выявления переведенных машиной предложений в тексте используется SVM-Light алгоритм для классификации.

Для каждого предложения рассчитываются следующие величины:

- энтропия языковой модели 3-грамм, обученной на 4,4 млн английских текстах (нужны для обнаружения локальных ошибок),

- аналитическая оценка модели 2 из (Collins, 1997), нормализованная по числу слов (более низкая величина для неоригинальных предложений),

- деревья происхождения слов;

- главное слово базового словосочетания и зависимости слов по частям речи.

Выявляются пять типов словесных зависимостей: субъект-глагол, глагол-объект, прилагательное-существительное, глагол-наречие и предлог-объект.

Сравниваются пары предложений "переведенный/оригинальный" текст.

Метод построения лингвистических моделей

Еще один способ [14] выявления машинного перевода относительно оригинального текста: выбирается целевой язык. Для любого другого языка составляется лингвистическая модель 1-,2-,3-, и 4-грамм из текстов, переведенных с этого языка на целевой, из текстов, переведенных на целевой язык с других языков и из оригинальных текстов целевого языка. Сравнивается perplexity: расстояние от модели до оригинальных текстов. Данные берутся из European Parliament Proceedings Parallel Corpus 1996-2011. Чем меньше perplexity, тем лучше модель.

Метод анализа частотных распределений

Следующий метод [10] позволяет выявить поисковой спам, автоматически порожденный с помощью цепей Маркова. Метод подразумевает анализ частотных распределений стилистических и жанровых особенностей текста. Идея метода – тексты,

сгенерированные на основе цепей Маркова, будут отличаться по определённым характеристикам от естественных текстов.

Каждая характеристика представляет собой положительное вещественное число. Группы сложно контролируемых автором текста характеристик:

1. Среднее количество символов в словах;
2. Среднее количество слогов в слове;
3. Доля слов длиннее 7 символов;
4. Доля слов более чем из 7 слогов;
5. Доля слов из слога;
6. Доля слов из двух слогов;
7. Минимальное количество слогов в одном предложении;
8. Максимальное количество слогов в одном предложении;
9. Количество частиц «бы»;
10. Количество частиц «ну», «вот», «ведь»;
11. Среднее количество знаков пунктуации на предложение;
12. Среднее количество знаков экспрессивной пунктуации («!», «?», «…»);
13. Среднее количество слов, начинающихся с заглавной буквы;
14. Доля различных частей речи:
    a. Доля глаголов среди слов;
    b. Доля прилагательных среди слов;
    c. Доля существительных среди слов;
    d. Доля числительных среди слов;
    e. Доля порядковых числительных среди слов;
    f. Доля наречий среди слов;
    g. Доля частиц среди слов;
    h. Доля предлогов среди слов;
    i. Доля частиц среди слов;
    j. Доля междометий среди слов;
15. Дисперсии количества различных частей (из п.15) речи по предложениям;
16. Доля местоимений первого лица;

17. Доля местоимений второго лица;

18. Доля глаголов по временам:

    a. Доля глаголов настоящего времени;

    b. Доля глаголов прошедшего времени;

    c. Доля глаголов не прошедшего времени;

19. Доля существительных по родам:

    a. Доля существительных мужского рода среди слов и среди существительных;

    b. Доля существительных женского рода среди слов и среди существительных;

    c. Доля существительных среднего рода среди слов и среди существительных;

20. Сжатие текста различными алгоритмами:

    a. bz2;

    b. zlib2;

21. Частотность употребления слов (оценка распределения по гистограмме частотностей).

Для определения текстов, созданных с помощью генераторов, выделенные характеристики объединяются в обучаемый классификатор, строящий зависимость между выделяемыми характеристиками и методом порождения документа.

Данные для эксперимента были получены из коллекции веб-страниц ROMIP (20 000 текстов) + тестируемая выборка из 10000*3 документов, порожденных различными генераторами текстов.

Алгоритмы обучения:

- SVMLight (на основе опорных векторов с использованием линейного ядра). F-мера - 71,69%.

Строит гиперплоскость, разделяющую разные классы объектов в пространстве признаков. Метод опорных векторов позволяет максимизировать зазор между классами, что способствует более качественной классификации.

Все параметры берутся по умолчанию. После процедуры обучения производится подбор порога классификации таким образом, чтобы сбалансировать точность и полноту классификатора на тренировочном наборе.

-Dtree (на основе деревьев решений). F-мера 91,57%

Основа – алгоритм построения деревьев решений C4.5 (выбор атрибута происходит на основании нормализованного прироста информации).

Каждое дерево – двоичное. Каждая вершина, не являющаяся листом, помечена номером признака и значением, по которому происходит разбиение набора документов на две части. Листы дерева помечены вероятностями принадлежности документа естественному или искусственно сгенерированному тексту. Дерево строится с корня. В начале, в корень дерева помещается часть тренировочного набора. Затем, в каждом листе выбирается такой признак и такое значение разбиения, которые минимизируют информационную энтропию в наборах, полученных после разбиения. В случае если энтропия в наборах, полученных после разбиения, меньше, чем в исходном наборе, для данного листа строится левые и правые поддеревья, и лист помечается номером соответствующего признака и порогом разбиения. Затем набор распределяется по левому и правому поддереву в соответствии с выбранным разбиением.

После построения дерева для каждого листа вычисляется вероятность того, что тексты, попавшие в этот лист, являются естественными или искусственными. Для этого документы распределяются по листам построенного дерева, затем для каждого листа вычисляется доли «хороших» и «плохих» документов, попавших в данный лист, которые и записывается в лист дерева. Чтобы минимизировать эффект переобучения на тренировочном наборе дерево строится на одной части тренировочного набора, а вероятности вычисляются по другой. При обучении по одному и тому же набору строится несколько деревьев решений. При построении каждого дерева тренировочный набор произвольным образом делится пополам. Первая половина используется для построения дерева, вторая используется для вычисления вероятностей принадлежности текста той или иной категории в каждом листе дерева. Деревья объединяются в один классификатор с помощью простой процедуры голосования. При классификации документа вычисляется, в какой лист он попадает в каждом дереве. После этого вычисляется сумма вероятностей принадлежности естественным и искусственным текстам по всем деревьям. Документу присваивается та метка, сумма вероятностей которой наибольшая.

Метод сравнения языковых моделей

Ещё один метод [7] по обнаружению спама использует сравнение языковых моделей. Метод можно использовать для 2-х случаев: для обнаружения комментариев, уже содержащих спам, и для блокировки нового входящего спама.

Для этого производится оценка методом максимального правдоподобия с JelinekMercer smoothing для построения корректной языковой модели с расстоянием Кульбака-Лейблера между разными текстами для оценки выбросов:

$$KL(\theta_1 \| \theta_2) = \sum_w p(w \mid \theta_1) \log \frac{p(w \mid \theta)_1}{p(w \mid \theta_2)} ,$$

где p(w|θi) — вероятность встретить слово w в соответствии с моделью θi.

То есть на имеющейся странице с комментариями «спам» и «не спам» вычисляется КЛ-расстояние для всех комментариев. Полученные величины можно рассматривать как извлеченные из распределения вероятностей, которые являются смесью распределений Гаусса: каждый гауссиан представляет отдельную лингвистическую модель. Гауссиан с наименьшей средней – наименьшее расстояние от лингвистической модели до оригинального сообщения – и есть модель языка, которая ближе всего к исходному сообщению.

Для оценки параметров используется EM-алгоритм. Наконец, комментарий расценивается как спам, если его КЛ-расстояние от сообщения в блоге скорее всего извлечено из спам-гауссиана, чем из правильного. Для определение этого рассчитывается пороговое значение.

Для исследования были взяты 50 случайных блог-постов и 1024 комментария к ним (как спам, так и не спам). В результате обнаружение спама происходит с 80% точностью (точность варьируется в зависимости от порогового множителя).

Метод анализа словосочетаний

Следующий метод [6] обнаружения фрагментов переведенного машиной текста в пространстве Интернета предполагает выявление в тексте фраз, согласующихся по «внутреннему» смыслу, но не согласующихся с другими фразами в предложении. В качестве данных были взяты англо-японские параллельные тексты.

Используются три параметра для обнаружения искусственного текста: "беглость", "грамматическая неточность" и "полнота" фраз в тексте.

Гладкость: Используются две независимых характеристики лингвистических моделей - fw и fw,MT. Обучение происходит на оригинальных текстах и на искусственных текстах. На вход поступает предложение и для него рассчитываются fw и fw,MT.

Грамматическая неточность: обучение происходит на последовательностях частей речи, созданных человеком и машиной. Соответственно вычисляются fpos и fpos,MT. Так же оцениваются комбинации фраз, которые нельзя покрыть диапазоном H-грамм. Для этого используются служебные слова, которые редко встречаются по одному, но часто встречаются в словосочетаниях. Кроме того, ищутся фразы, состоящие из двух "подфраз" (пр. "not only but also") - такие фразы характерны для оригинальных текстов - оценивается вероятность появления второй "подфразы" при появлении "первой".

Полнота: Рассчитывается величина $f_c(s) = \sum_{i \in k} w_i \delta(i,s)$ где wi - вес i-й фразы, δ(i, s) -

дельта-функция Кронекера (=1, если предложение s содержит i-ю фразу; иначе 0). Вес задаётся важностью фразы.

Также вводится flen как характеристика длины предложения для того, чтобы избежать смещения лингвистической модели на основе коротких предложений.

Таким образом строится вектор f(fw , fw,MT...flen), каждое f нормализируется: N(0,1). Далее используется метод опорных векторов для классификации принадлежности текста.

Результат работы метода: 95,8% точности для отдельных предложений и 80.6% для зашумлённых веб-страниц.

Метод классификации по неконтролируемым автором признакам

Обнаружение спама путём проведения содержательного анализа предлагается статье [13]. Decision Tree классификация на спам/не-спам использует в качестве признаков сжимаемость документа, среднюю вероятность триграмм, долю частых слов в документе и т.д.. Данные для исследования - 105, 484, 446 веб-страниц. Полученная таким способом точность выявления спама - 86.2%.

Признаки спама:

- спам-страницы содержат бОльшее количество редко используемых в обычной речи слов.
- формулировки поисковых запросов присутствуют в заголовках спам-страниц, причем чрезмерное количество слов в названии страницы является лучшим показателем спама, чем количество слов в содержании странице.
- средняя длина спам-слова – 8-10 символов.
- спам-страницы могут повторять в своем содержимом несколько раз одну и ту же информацию для повышения ранга страницы в поисковике. Обнаружение выявляется методом шилинга (последовательностей фиксированной длины k, состоящих из соседних слов):

Для каждой цепочки вычисляются 84 «дактилограммы» по алгоритму Рабина — Карпа с помощью взаимно-однозначных и независимых функций, использующих случайные наборы («min-wise independent») простых полиномов. В результате каждый документ представлялся 84 шинглами, минимизирующими значение соответствующей функции. Затем 84 шингла разбиваются на 6 групп по 14 (независимых) шинглов в каждой (супершинглы). Если два документа имеют сходство, например, p ~ 0.95 (95%), то 2 соответствующих супершингла в них совпадают с вероятностью p14 ~ 0.9514 ~ 0.49 (49%). Поскольку каждый документ представляется 6 супершинглами, то вероятность

того, что у двух документов совпадут не менее 2-х супершинглов, равна: 1 - (1-0.49)6 - 6•0.49•(1-0.49)5 ~ 0.90 (90%).

Таким образом, для эффективной проверки совпадения не менее 2-х супершинглов каждый документ представляется всевозможными попарными сочетаниями из 6 супершинглов («мегашинглов»). Число таких мегашинглов равно 15 (число сочетаний из 6 по 2). Два документа сходны по содержанию, если у них совпадает хотя бы один мегашингл. При обнаружении повторов содержание веб-страницы сжимается.

Слова могу быть рандомно взяты из словаря в соответствии с распределением частот слов в языке. Для анализа содержания страницы на грамматическую и семантическую корректность используется поиск вероятностной локальной последовательности: каждый документ сегментируется из большого корпуса в n-граммы n последовательных слов (3-4). Определяется вероятность n-грамм wi+1...wi+n начиная с i+1 слова.

N-граммы выбираются независимо друг от друга. Вероятность документа с k n-граммами является произведением индивидуальных вероятностей. Длина документа нормализируется.

Документы, состоящие из часто встречающихся n-грамм вероятнее всего являются спамом, а также документы, составленные из невероятных n-грамм. Далее создаётся классификатор, учитывающий особенности проверяемой страницы для определения ее принадлежности спам/не спам.

Для выбора параметров для идентификации искусственно созданных текстов предлагается провести :

− 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.

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

Данными для исследования служат сгенерированные машиной тексты с помощью SyMonym и Article Clone Easy.

Литература:

[1] *Grechnikov E.A., Gusev G.G., Kustarev A.A., Raigorodsky A.M.*: Detection of Artificial Texts, Digital Libraries: Advanced Methods and Technologies, Digital Collections: Proceedings of the XI All-Russian Research Conference RCDL'2009. Petrozavodsk: KRC RAS, 2009. Pp. 306-308.

[2] *Simon Corston-Oliver, Michael Gamon and Chris Brockett*: A machine learning approach to the automatic evaluation of machine translation, Proceeding ACL '01 Proceedings of the 39th Annual Meeting on Association for Computational Linguistics, Pages 148-155

[3] *Tanguy Urvoy, Thomas Lavergne, Pascal Filo-che*: Tracking Web Spam with Hidden Style Similarity. AIRWEB'06, August 10, 2006, Seattle, Washington, USA.

[4] *Thomas Lavergne, Tanguy Urvoy, Francois Yvon*: Filtering artificial texts with statistical machine learning Techniques. Language Resources and Evaluation, March 2011, Volume 45, Issue 1, pp 25-43

[5] *Witten IH, Frank E: Data Mining*: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann Publishers (2011)

[6] *Yuki Arase, Ming Zhou*: Machine Translation Detection from Monolingual Web-Text. Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 1597–1607, Sofia, Bulgaria, August 4-9 2013

[7] *Baayen R.H.*: Word Frequency Distributions. Kluwer Academic Publishers, Amsterdam, The Netherlands, (2001)

[8] *Clarkson, P. and R. Rosenfeld*. Statistical Language Modeling Using the CMU-Cambridge Toolkit. Proceedings of Eurospeech97. 2707- 2710.

[9] *Chickering, D. M., D. Heckerman, and C. Meek.* A Bayesian approach to learning Bayesian networks with local structure. In Geiger, D. and P. Punadlik Shenoy (Eds.), Uncertainty in Artificial Intelligence: Proceedings of the Thirteenth Conference. 80-89

[10] *Vladimir N. Vapnik.* 1995. The nature of statistical learning theory. Springer.

[11] *Chen SF, Goodman JT* (1996) An empirical study of smoothing techniques for language modeling. In: Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics (ACL), Santa Cruz, pp 310–318

[12] *Honore A* (1979) Some simple measures of rich-ness of vocabulary. In: Association for Literary and Linguistic ´Computing Bulletin, vol 7(2), pp 172–177

[13] *Sichel H* (1975) On a distribution law for word frequencies. In: Journal of the American Statistical Association, vol 70, pp 542–547

[14] *Thomas Lavergne, Tanguy Urvoy, Francois Yvon*: Detecting Fake Content with Relative

entropy Scoring, PAN 2008

[15] *K. Seymore and R. Rosenfeld*: Scalable backoff language models, In ICSLP '96, volume 1, pp. 232–235, Philadelphia, PA, (1996).

[16] *A. Stolcke*. Entropy-based pruning of backoff language models, 1998

[17] C. *D. Manning and H. Schutze*: Foundations of Statistical Natural Language Processing, The MIT Press, Cambridge, MA, 1999.