УДК 004.931

АНАЛИЗ РАЗВИТИЯ ВЕРОЯТНОСТНЫХ МОДЕЛЕЙ ЯЗЫКА В ЕСТЕСТВЕННО-ЯЗЫКОВЫХ ПРИЛОЖЕНИЯХ

С.Ю. Мельников
Кафедра теории кибербезопасности института компьютерных наук и телекоммуникаций факультета физико-математических и естественных наук Российского университета дружбы народов имени Патриса Лумумбы
Россия, 117198, г. Москва, ул. Миклухо-Маклая, 6
E-mail: melnikov-syu@rudn.ru

В.А. Пересыпкин
Академия Криптографии РФ
Россия, 119331, г. Москва, пр-кт Вернадского, 12
E-mail: info@cryptoacademy.gov.ru

Ключевые слова: вероятностная модель языка, нейросетевая модель языка, энтропия, перплексия, распознавание речи, OCR, коррекция текста.

Аннотация: В работе проведен анализ развития вероятностных моделей языка, предпринята попытка связать их с развитием персональных вычислительных средств и математического аппарата моделирования. Рассмотрены наиболее востребованные, на сегодняшний день, статистические и нейросетевые модели языка, которые активно используются на практике в естественно-языковых приложениях, связанных с задачами распознавания (распознавание речи, OCR, коррекция искаженных текстов и др.). Отмечено, что развитие технологий языкового моделирования привело к существенному снижению перплексии и энтропии разрабатываемых моделей, что, в свою очередь, позволило существенно повысить эффективность распознавания и коррекции.

1. Введение

Интерес человеческого общества к изучению особенностей языка общения, характеристик созданных на нем текстов возник достаточно давно. Одной из первых математических работ в этой области является статья А.А. Маркова [1] с анализом частотных свойств цепочек символов, составляющих текст «Евгения Онегина». Эта статья считается родоначальницей теории цепей Маркова.

Продолжением исследований в области языка стали попытки построения языковых моделей, отображающих часть языковой реальности с ее важными свойствами, но более простых, чем эта реальность. Одними из наиболее востребованных моделей языка (текста) являются так называемые вероятностные модели (используется также выражение «статистические модели языка», statistical language models, SLM), которые позволяют оценивать правдоподобие фрагмента текста, вычислять степень его согласия с основными статистическими характеристиками языка. Их основная задача – предсказать следующий элемент текста (символ, морфему или слово), если известны предшествующие элементы. Чем точнее это предсказание, тем лучше модель.

Создаваемые модели языка постепенно приближаются по своим свойствам к объекту исследований – естественному человеческому языку. Двигателем этого процесса служит постоянное расширение областей практического использования

языковых моделей и возрастающие требования к их точности. Основанием для развития моделей языка служат, во-первых, несовершенство текущего математического аппарата моделирования, во-вторых, недостаток имеющихся исходных данных для моделирования, в-третьих, ограничения по доступным вычислительным ресурсам.

2. Области применения

До появления в 50-60-х годах XX века первых вычислительных средств использовались построенные вручную модели языка на символах или словах, в которых трудоемкие задачи ложились на плечи экспертов-лингвистов. С появлением в конце 90-х годов XX века производительных персональных вычислительных средств и существенным расширением базы доступных лингвистических ресурсов расширяется и круг практических приложений языковых моделей.

Криптография. Одной из первых практически важных областей, в которых возникла необходимость использования вероятностных моделей языка, стала криптография. Количественное описание статистических свойств текста необходимо для расчета стойкости шифров. Такие описания были получены в работах К. Шеннона [2] и В.Котельникова [3].

Текстовая стеганография. Текстовая стеганография сегодня является одним из актуальных и многообещающих научных направлений в области информационной безопасности. Методы сокрытия информации в естественном тексте условно можно разделить на два класса: редактирование и генерация.

В методах, основанных на редактировании текста, таких, как замена синонимов, модификация заголовков, расстановка пробелов и др., способ модификации текста является стеганографическим контейнером. Оценка информационной емкости таких контейнеров требует использования языковых моделей [4].

В генеративных методах (Generative Linguistic Steganography, GLS) в зависимости от содержимого контейнера генерируются различные блоки текста, такие системы позволяют встраивать большие объемы информации [5].

Распознавание речи. Одной из основных составных частей систем распознавания речи [6] является языковая модель. Классические системы распознавания содержат две независимые модели: акустическую и языковую. С помощью этих моделей в ходе распознавания оцениваются строящиеся гипотезы о произнесенном фрагменте речи. Сама языковая модель обучается [7] на больших текстовых корпусах, стиль текстов которых близок к стилю распознаваемых.

Точность системы распознавания речи во многом определяется точностью использованной модели языка. Этот факт послужил мощным стимулом для развития техники языкового моделирования, было предложено множество усовершенствований базовых N-граммных моделей текста, а наиболее значимым шагом стали нейросетевые модели [8]. Возрастающее распространение систем распознавания речи требует как повышения точности используемых моделей, так и возможности реализации на вычислительных платформах с ограниченными ресурсами (например, мобильных устройствах) [9].

Оптическое распознавание символов (OCR системы). Языковые модели являются необходимым дополнением для получения точных результатов в системах оптического распознавания символов, хотя при распознавании документов хорошего качества роль языковых моделей не столь значительна. Существуют изображения, текст на которых сильно отличается от общей языковой модели (чеки, рецепты и другие специализированные классы документов). Для этих условий предложены [10]

методы создания и присоединения языковой модели на основе слов для конкретной предметной области к общей языковой модели в системе OCR.

Коррекция ошибок в тексте. Программные средства коррекции текстов на распространенных языках уверенно исправляют тексты с малым числом искажений. Однако в случае текстов с высоким уровнем искажений, вне зависимости от их происхождения (набранных с ошибками на клавиатуре, полученных в результате распознавания речи в условиях шумов и др.), такие средства показывают неудовлетворительные результаты [11].

Как правило, для коррекции искажений используются языковые модели, которые позволяют строить цепочки словоформ скорректированного текста из колонок вариантов слов для каждого искаженного фрагмента текста [12].

Машинный перевод. Языковые модели, используемые в системах машинного перевода, при формировании следующего фрагмента перевода используют как само переводимое предложение, так и результаты выполненного предыдущего частичного перевода. В настоящее время преобладают модели машинного перевода E2E. Языковая модель неявно обучается при обучении самой системы E2E. В последнее время лучшие результаты в машинном переводе получены моделями на основе нейросетевой архитектуры трансформеров [13].

3. Критерии эффективности моделей языка

Для оценки эффективности языковых моделей обычно используется правдоподобие новых данных, которые не участвовали в обучении модели. Среднее логарифмическое правдоподобие (average log likelihood) новых данных определяется следующим образом:

$$Average\ -Log-Likelihood(D|M) = \frac{1}{N}\sum_{i=1}^{N} \log P_M(D_i),$$

где $D=\{D_1, D_2, \ldots, D_N\}$ – это новые данные, M – используемая языковая модель.

Последняя величина также может рассматриваться как эмпирическая оценка кросс-энтропии (cross-entropy) истинного (но неизвестного) распределения P с учетом распределения модели P_M:

$$cross-entropy(P;\ P_M) = -\sum_D P(D) \cdot \log P_M\,(D).$$

На практике эффективность языковой модели обычно измеряется с помощью перплексии (perplexity):

$$perplexity(P;\ P_M) = 2^{cross-entropy(P;\ P_M)}.$$

Перплексия может интерпретироваться как средний геометрический коэффициент ветвления (branching factor) языка в соответствии с моделью. Перплексия является функцией как языка, так и модели. С помощью перплексии можно оценивать насколько хороша модель – чем лучше модель, тем ниже перплексия.

4. Статистические модели языка

Символьная модель $\mathbf{M_1}$. В рамках данной модели источник сообщений вырабатывает знаки текста по схеме независимых испытаний согласно вероятностному распределению, заданному на знаках алфавита A. Вероятность порождения таким источником последовательности знаков $a_{i_1}\ldots a_{i_L}$ равна

$$p\left(a_{i_1}\ldots a_{i_L}\right) = \prod_{l=1}^{L} p_{i_l}$$

Энтропия модели

$$H_M = -\sum_{i=1}^{z} p_i \log p_i$$

Данная модель является наиболее грубой, но она очень проста в реализации, вплоть до ее построения экспертом-лингвистом вручную. Значение энтропии велико (например, ~ 4.08 бит/символ для английского языка), что ограничивает сферу её применения лишь случаями, когда имеются существенные ограничения на сложность и вычислительный ресурс.

Символьная Марковская модель $\mathbf{M_r}$ конечного порядка r. В моделях данного класса источник моделируется однородной цепью Маркова порядка r (r = 2,3,4,5…), поэтому в обиходе используется термин «r - граммная марковская модель». Модель порядка r задается с помощью вектора начальных вероятностей $\vec{p} = (p_{i_1 \dots i_{r-1}})$ размерности z^{r-1}, где $z = |A|$, задающим начальное распределение на последовательностях $a_{i_1} \dots a_{i_{r-1}}$ длины $r-1$ символов алфавита A, и матрицей переходных вероятностей размерности $z^{r-1} \times z$, состоящей из условных вероятностей появления символа a_{i_r} после последовательности символов $a_{i_1} \dots a_{i_{r-1}}$.

Вероятность порождения таким источником последовательности знаков $a_{i_1} \dots a_{i_L}$ равна

$$p(a_{i_1} \dots a_{i_L}) = p_{i_1 \dots i_{r-1}} \cdot \prod_{l=r}^{L} p\left({a_{i_l}}/{a_{i_{l-r+1}} \dots a_{i_{l-1}}} \right).$$

Энтропия такой модели есть величина

$$H^{(r)} = -\sum_{(a_{i_1} \dots a_{i_r}) \in A^r} p(a_{i_1} \dots a_{i_r}) \cdot \log p\left({a_{i_r}}/{a_{i_1} \dots a_{i_{r-1}}} \right).$$

Значение энтропии заметно снижается (для биграммной модели ~3.23 бит/символ, для триграммной ~2.83 бит/символ и т.д.) по сравнению с символьной моделью $\mathbf{M_1}$.

Марковская модель порядка $\boldsymbol{r}$ имеет $\boldsymbol{z^{r-1}}$ состояний, что до появления достаточно производительных персональных вычислительных средств вносило ограничения на порядок модели или предполагало обработку на специализированных вычислителях. Это было связано с объемом оперативной памяти, необходимой для реализации в ней как самой модели, так и рабочих массивов целевого алгоритма.

Для построения модели с большим числом состояний необходимо иметь обучающую выборку, объем которой сопоставим с числом состояний модели. На практике считается достаточным объем выборки, превышающий число состояний модели в 5-10 раз**.** Для создания биграммной и триграммной символьной модели языка речь шла об обучении на корпусах текстов объемом единицы и до нескольких десятков Мбайт символов текста, для четырех и пятиграммной символьной модели языка необходимы корпуса в десятки и до нескольких сотен Мбайт символов текста.

Словарная Марковская модель $\mathbf{Ms_r}$ конечного порядка r. Строится аналогично символьной Марковской модели M_r, только вместо символов в качестве текстовых единиц используются слова из обучающего корпуса текстов. По сравнению с символьной моделью значение энтропии существенно снижается уже для словарной модели первого порядка (~3.0 бит/символ).

Необходимо иметь значительный объем обучающей выборки для построения достаточно представительной модели. Эксперименты со словарной моделью первого порядка показывают, что стабилизация значения энтропии модели происходит тогда, когда объём обучающего корпуса текстов составляет нескольких сотен Мбайт.

Марковские VLMM-модели текста конечного порядка r. В качестве вероятностной модели текста на естественном языке можно также применять Марковские VLMM-модели текста с переменной длиной зависимости (variable length Markov models, VLMM), позволяющие достаточно эффективно аппроксимировать *r*-граммные языковые модели со сглаживанием, но требующие меньшего объема памяти. Энтропия VLMM-моделей текста может быть существенно снижена (до 1-1.2 бит/знак), что позволяет в существенно повысить точность распознавания (с 60-70% до 80-90% единиц текста).

5. Нейросетевые модели языка

С ростом вычислительной мощности современных параллельных вычислителей нейросетевые модели стали широко применяться в области моделирования языка. Нейросетевые модели языка основаны на непрерывных векторных представлениях слов. Если в N-граммных словарных моделях слово полностью определяется целым числом – своим индексом в словаре, то векторные, многомерные представления позволяют учитывать разные и не всегда очевидные меры схожести между словами, в том числе семантические, контекстные, морфологические свойства. Такое представление напоминает модели на классах слов, но позволяет учитывать множество способов разбиения на классы одновременно, что существенно улучшает качество моделей языка.

Рекуррентные сети – сети, в которых помимо связей, передающих данные в прямом порядке, есть связи, передающие данные на вход предыдущего слоя, т.е. реализуется обратная связь, добавляющая зависимость результатов от данных, поданных ранее. Рекуррентные свойства модели могут быть реализованы добавлением связи между выходом слоя и предшествующим слоем, либо применением специальных видов скрытых слоев как Long Short-Term Memory (LSTM) [14].

Последние годы лучшие результаты показывают нейросетевые модели, основанные на так называемых трансформерах. Трансформер (transformer) - архитектура глубоких нейронных сетей, основанная на механизме внимания и не использующая рекуррентные нейронные сети.

В Табл. 1 приведены достигнутые значения перплексий разных моделей, полученные на корпусе [15] англоязычных текстов, содержащем миллиард слов.

Таблица 1. Значения перплексии для разных моделей.

Модель	Перплексия
5-граммная статистическая модель со сглаживанием Кнессера-Нея [16]	67.6
LSTM [14]	30.0
Трансформер [17]	23.0

В [18] отмечается, что, несмотря на то, что, в целом, нейросетевые модели оказываются точнее N-граммных, в некоторых случаях, например, для ограниченного обучающего корпуса, N-граммные модели являются более предпочтительными. К таким случаям могут относиться корпуса текстов, собранные на т.н. «малоресурсных» языках, к которым, в частности, относятся многие национальные языки нашей страны.

6. Заключение

В работе рассмотрены наиболее востребованные, на сегодняшний день, статистические и нейросетевые модели языка, которые активно используются на практике в естественно-языковых приложениях.

Развитие технологий языкового моделирования в последние десятилетия привело к существенному снижению перплексии и энтропии разрабатываемых моделей, что позволяет в естественно-языковых приложениях достичь высоких значений показателей точности распознавания, сравнимых с возможностями человека. Отмечено, что новые модели для своего обучения требуют больших объемов обучающих корпусов и значительных вычислительных ресурсов.

Проведено сравнение показателей эффективности наиболее предпочтительных для моделирования статистической и нейросетевой моделей языка при различных условиях моделирования. Сделан вывод, что при общей тенденции предпочтительности нейросетевой модели языка, в ряде случаев статистическая модель может являться более выигрышной.

Список литературы

1. Марков А.А. Пример статистического исследования над текстом «Евгения Онегина», иллюстрирующий связь испытаний в цепь // Известия Имп. Акад. наук, серия VI, Т. X, № 3, 1913. С. 153-162
2. Shannon C. Communication theory of secrecy systems // Bell Sys. Tech. J. 1949. Vol. 28. P. 656-715.
3. Андреев Н.Н., Петерсон А.П., Прянишников К. В., Старовойтов А.В. Основоположник отечественной засекреченной телефонной связи // Радиотехника. 1998. № 8. С. 8-12.
4. Lingyun Xiang, Yan Li, Wei Hao, Peng Yang, and Xiaobo Shen. Reversible natural language watermarking using synonym substitution and arithmetic coding // Comput., Mater. Continua. 2018. Vol. 55, No. 3. P. 541-559.
5. Xiang L., et al. Generative Linguistic Steganography: A Comprehensive Review // KSII Transactions on Internet & Information Systems. 2022. Vol. 16, No. 3.
6. Karpagavalli S., Chandra E. A review on automatic speech recognition architecture and approaches // International Journal of Signal Processing, Image Processing and Pattern Recognition. 2016. Vol. 9, No. 4. P. 393-404.
7. Кипяткова И. С., Карпов А. А. Разработка и исследование статистической модели русского языка // Информатика и автоматизация. 2010. №. 12. С. 35-49
8. Чучупал В.Я. Нейросетевые модели языка для систем распознавания речи // Речевые технологии / Speech Technologies. 2020. №. 1-2. С. 27-47.
9. Чучупал В.Я. Способы уменьшения вычислительной сложности нейросетевых языковых моделей // Речевые технологии / Speech Technologies. 2020. №. 3-4. С. 16-29.
10. Garst P., Ingle R., Fujii Y. OCR Language Models with Custom Vocabularies // International Conference on Document Analysis and Recognition // Cham: Springer, 2023. P. 101-115.
11. Бирин Д.А., Мельников С.Ю., Пересыпкин В.А., Писарев И.А., Цопкало Н.Н. Об эффективности средств коррекции искаженных текстов в зависимости от характера искажений // Известия ЮФУ. Технические науки. 2018. № 8 (202). С. 104-114. DOI: 10.23683/2311-3103-2018-8-104-114. EDN SXENWY.
12. Вахлаков Д.В., Германович А.В., Мельников С.Ю., Пересыпкин В.А., Цопкало Н.Н. О точности и трудоемкости многоэтапного метода коррекции искаженных текстов в зависимости от степени искажения // Известия ЮФУ. Технические науки. 2021. №. 7 (224). С. 130-142. DOI: 10.18522/2311-3103-2021-7-130-142. EDN BFQZHT.
13. Qiang Wang, Bei Li, Tong Xiao, Jingbo Zhu, Changliang Li, Derek F Wong, and Lidia S Chao. Learning deep transformer models for machine translation // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019. P. 1810-1822.
14. Gers F.A., Schmidhuber J., Cummins F. Learning to forget: Continual prediction with LSTM // Neural computation. 2000. Vol. 12, No. 10. P. 2451-2471.
15. Chelba C., et al. One billion word benchmark for measuring progress in statistical language modeling. arXiv preprint arXiv:1312.3005. 2013.
16. Chen W., Grangier D., Auli M. Strategies for training large vocabulary neural language models. arXiv preprint arXiv:1512.04906. 2015.

17. Baevski A., Auli M. Adaptive input representations for neural language modeling. arXiv preprint arXiv:1809.10853. 2018.
18. Мельников С.Ю., Пересыпкин В.А. Об эволюции классических вероятностных моделей языка в естественно-языковых приложениях // Вестник современных цифровых технологий. 2023. № 16. С. 4-14. EDN YDIGDT.