histories go through a procedure for processing big data by the Hadoop cluster at NovSU. The result of processing will be a list of links to http pages with relevant, interesting and high-quality publications on the subject under study.

Keywords: information technology in education, big data, expert systems, personalization of education

References

1. Bot Code Examples URL: https://core.telegram.org/bots/samples (Accessed: 29.11.2017).
2. Marinescu, D. C. (2013) Cloud Computing Theory and Practice. USA: Morgan Kaufmann is an imprint of Elsevier. 415 p.

УДК 004.912, 51-74

СТАТИСТИЧЕСКИЕ МОДЕЛИ ЯЗЫКА

Мурат Казбекович Датиев
аспирант
г. Владикавказ
Московский институт радиотехники, электроники и автоматики

Александр Юрьевич Кулай
старший научный сотрудник
г. Владикавказ
Московский институт радиотехники, электроники и автоматики

Казбек Муратович Датиев
к.т.н., профессор
datiev_skgmi@mail.ru
г. Владикавказ
Северо-Кавказский горно-металлургический институт

Аннотация. В работе рассматриваются современные статистические модели языка. Дано определение применяемым критериям эффективности моделей. Описываются следующие статистические модели языка: n-граммные модели, модели на основе деревьев решений, лингвистически мотивированные модели.

Ключевые слова: статистические языковые модели, n-граммные модели, перплексия.

Введение. При помощи статистического моделирования языка (Statistical Language Modeling (SLM)) формализуются закономерности естественного языка для улучшения производительности различных естественно-языковых приложений. Статистическое языковое моделирование строит оценку вероятностного распределения различных лингвистических единиц, таких как слова, предложения, тексты [Розенфельд, 1996]. Статистические модели языка применяются в большом количестве приложений: распознавание речи, машинный перевод, классификация документов, информационный поиск, распознавание рукописного текста, проверка орфографии и др.

SLM использует технологии статистической оценки, применяя для обучения языковые данные – текст. Из-за категориальной природы языка и огромного словарного запаса статистические технологии должны оценивать большое количество параметров и, следовательно, сильно зависят от доступности значительного количества обучающих

данных. За последние десятилетия большое количество текстов различных типов стало доступно в сети Интернет. В результате, в областях, где подобные данные стали доступными, качество языковых моделей резко улучшилось. Однако, сейчас это усовершенствование приближается к своему пределу. Даже если онлайн-тексты накапливаются по экспоненте, то качество статистических моделей языка, используемых в настоящее время, навряд ли значительно улучшится.

Как это ни удивительно, но самые производительные SLM технологии используют очень мало знаний о языке как таковом. Например, *n*-граммные модели никак не используют информацию, что моделируется язык, они аналогично могут моделировать последовательности произвольных символов, за которыми не стоит никакой глубокой структуры или смысла. Предпринимаются различные попытки создать модели языка, которые применяют знания о том, что на самом деле представляет собой язык. По этому поводу один из первых инициаторов статистического подхода к языковому моделированию Фред Джелинек сказал: «необходимо вернуть язык в языковое моделирование». Было предпринято определенное количество попыток включить лингвистическую структуру, теории или знание в статистические модели языка, большинство этих попыток было мало результативным.

Применение моделей языка. Статистическое языковое моделирование – это вероятностное распределение *P(s)* для всех возможных предложений *s* (или слов, или устных высказываний, документов или любых других лингвистических единиц).

Сравним статистическое моделирование языка с вычислительной лингвистикой (computational linguistics). Следует отметить, что обе эти области имеют нечеткие границы и значительно совпадают. Пусть *S* – последовательность слов данного предложения, а *H* – какая-то связанная с ним скрытая структура (т.е. дерево грамматического разбора, смыслы слов и т.д.). Статистическое моделирование языка преимущественно нацелено на оценку вероятности *P(S)*, тогда как вычислительная лингвистика – больше на оценку условной вероятности *P(H|S)*. Естественно, если можно оценить совместную вероятность *P(S, H)*, то из него могут быть получены и *P(S)*, и *P(H|S)*. На практике это обычно невозможно. Статистические модели языка обычно используются в контексте байесовского классификатора, где они могут выполнять роль априорной функции или функции правдоподобия. Например, при распознавании речи исследуется акустический сигнал *a*, цель – найти предложение *s*, которое было произнесено с наибольшей вероятностью. При использовании байесовского подхода решение будет следующим:

$$s^* = \arg\max_s P(s \mid a) = \arg\max_s P(a \mid s) \cdot P(s),$$

здесь языковая модель *P(s)* играет роль априорной функции.

Однако, при классификации документов имеется документ *d*, цель – найти класс *c*, которому он принадлежит. Как правило, имеются образцы документов для каждого из *k* классов, из которых строятся *k* различных языковых моделей $\{P_1(d), P_2(d), \ldots, P_k(d)\}$. Используя байесовский классификатор, решение *c** следующее:

$$c^* = \arg\max_c P(c \mid d) = \arg\max_c P(d \mid c) \cdot P(c), \qquad (*)$$

здесь модель языка $P_c(d)$ исполняет роль функции правдоподобия.

Аналогично можно вывести роль моделей языка в байесовском классификаторе для других языковых технологий.

Критерии эффективности моделей языка. Для оценки качества техники моделирования языка обычно используется правдоподобие новых данных. Среднее

логарифмическое правдоподобие (average log likelihood) нового произвольного образца определяется следующим образом: $Average–Log–Likelihood(D \mid M) = \frac{1}{n}\sum_i \log P_M(D_i)$,

где $D=\{D_1, D_2, \ldots, D_N\}$ – это новый образец данных, M – используемая языковая модель. Последняя величина также может рассматриваться, как эмпирическая оценка кросс-энтропии (перекрестной энтропии (cross-entropy)) истинного (но неизвестного) распределения P с учетом распределения модели P_M:

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

Фактическая производительность языковой модели обычно определяется с помощью перплексии (perplexity): $perplexity(P; P_M) = 2^{cross–entropy(P; P_M)}$.

Перплексию может интерпретировать как средний (геометрический) коэффициент ветвления языка в соответствии с моделью. Это функция, как языка, так и модели. Применительно к функции модели, она оценивает, насколько хороша модель (чем лучше модель, тем ниже перплексия). Применительно к языку, она оценивает энтропию или сложность этого языка. В конечном счете, качество языковой модели должно быть определено её производительностью в конкретном приложении, для которого она разрабатывалась, а именно, процентом ошибок в данном приложении. Однако, процент ошибок, как правило, нелинейные и плохо понятные функции языковой модели. Более низкая перплексия обычно приводит к более низкому проценту ошибок, но в литературе встречается много обратных примеров. Эмпирически, уменьшение перплексии на 5% фактически незаметно, уменьшение на 10%-20% заметно и обычно (но не всегда) приводит к некоторому улучшению производительности приложения, уменьшение перплексии на 30% и выше является весьма существенным (и редким).

Некоторые недостатки языковых моделей. Даже самая простая языковая модель существенно влияет на приложения, в которых она используется (в этом можно убедиться, например, убрав языковую модель из системы распознавания речи). Однако, даже современные методы языкового моделирования далеко не оптимальны. Есть несколько причин:

Уязвимость в предметной области. Современные модели языка крайне чувствительны к изменениям в стиле, теме или жанре текста, на котором они обучались.

Ложное предположение о независимости. Для того чтобы оставаться «легко поддающимися обработке» фактически все используемые языковые методы моделирования принимают некоторую форму независимости между различными частями одного и того же документа.

Эксперименты по Шеннону. Клод Шеннон применял технику извлечения знания человека о языке, просив людей предсказывать следующий элемент текста. Шеннон использовал этот метод для оценки энтропии английского языка. Общее наблюдение во время подобных экспериментов заключается в том, что люди улучшают работу языковой модели достаточно легко и существенно. Они, вероятно, делают это за счет использования рассуждений на уровнях предметной области, лингвистики и здравого смысла.

Обзор основных техник статистического языкового моделирования. Практически все модели языка раскладывают вероятность предложения в произведение условных вероятностей: $P(s) \overset{def}{=} P(w_1 \dots w_n) = \prod_{i=1}^{n} P(w_i | h_i)$,

где w_i – i-ое слово в предложении, а $h_i = \{w_1, w_2, \dots, w_{i-1}\}$ - его история.

В настоящее время, помимо моделей на нейронных сетях, применяются следующие статистические модели языка:

- n-граммные модели;
- модели на основе деревьев решений;
- лингвистически мотивированные модели:
  - контекстно-независимая грамматика;
  - грамматика связей.

N-граммные модели (*n*-grams). n-граммные модели являются основными моделями языка в современных технологиях распознавания речи. Практически все коммерческие продукты распознавания речи используют некоторую форму n-граммных моделей. n-граммная модель понижает размерность задачи, моделируя язык как дискретный Марковский источник порядка *(n-1):* $P(w_i \mid h_i) \approx P(w_i \mid w_{i-n+1}, \dots, w_{i-1})$.

Значение n изменяет соотношение стабильности оценки (т.е. дисперсии) и ее правильности (т.е. смещения). Получение вероятностей используемых n-грамм до сих пор сталкивается с «проблемой нулевых переходов», даже на очень больших корпусах данных. Более того, даже среди найденных n-грамм, подавляющее большинство встречается лишь однажды, а многие другие характеризуются низкой встречаемостью. Поэтому непосредственное использование оценок максимального правдоподобия (ОМП) для нахождения оценок вероятностей n-грамм по их частотам нежелательна. Для решения данной проблемы предложены различные методы сглаживания. Один из перспективных подходов был предложен авторами данной работы в [Датиев, Кулай, Датиев, 2013].

Модели на основе деревьев решений (decision tree models). Деревья решений и алгоритмы типа CART впервые описаны в [Баль, Браун, де Соуза, Мерсер, 1989]. Двоичное дерево решений состоит из множества неконечных и конечных вершин. Каждая неконечная вершина связана с двоичным вопросом и имеет два перехода в вершины следующего уровня. Пример двоичного вопроса: «$w_{i-3} \in \{[and], [or], [not]\}$?». Подобным образом двоичное дерево решений произвольно разбивает пространство историй путем задавания бинарных вопросов об истории h на каждом из внутренних узлов. Каждая конечная вершина (лист) помечена вероятностным распределением для последующего слова $P(w|h)$, для оценивания которого применяются обучающие данные. Для снижения дисперсии оценки распределения в листе интерполируются с распределениями во внутренних узлах, встречающихся на пути от корня. Для того чтобы найти вероятность слова w_i с предысторией h_i необходимо пройти из корня по неконечным вершинам графа по пути, определяемым ответами на двоичные вопросы, пока не найдется конечная вершина. Вероятность слова w_i находится из распределения, которым помечен этот лист.

Лингвистически мотивированные модели (linguistically motivated models). Так как SLM основанны на интуитивном подходе к языку, в большинстве моделей лингвистическое содержание незначительно. Однако, некоторые технологии заимствованы непосредственно из грамматик, широко используемых лингвистами.

Контекстно-независимая грамматика (context free grammar). Контекстно-независимая грамматика [Чен, 1996] – грубая, но очень ясная модель естественного языка. Она состоит из словаря, набора нетерминальных символов и набора правил перехода и образования. Предложения формируются, начиная с начального нетерминального символа, посредством повторяющегося применения правил перехода, каждое из которых преобразует нетерминальный символ в последовательность терминальных (т.е. слов) и нетерминальных, пока не получится предложение, состоящее только из терминальных последовательностей.

Грамматика состоит из правил, которые определяют, как структуры одного уровня языка переходят в формы структур следующего уровня.

Грамматика связей (link grammar). Грамматика связей – модель, предложенная в [Слеатор, Тамперлей, 1991]. От остальных контекстно-независимых грамматик, грамматику связей отличает отсутствие явных составляющих и высокая степень лексикализации. Последнее свойство делает грамматику связей привлекательной при вероятностном моделировании.

Заключение. В работе описаны статистические модели языка, применяемые в настоящее время. Благодаря возросшему доступу к большим объемам данных, статистические модели показывают все более впечатляющие результаты. Несмотря на это, перед статистическими моделями языка по-прежнему стоит ряд сложностей, которые исследуются и решаются иностранными и отечественными авторами.

Список литературы

1. Баль Л.Р., Браун П.Ф., де Соуза П.В., Мерсер Р.Л. (1989) Статистическая языковая модель для распознавания речи, основанная на деревьях. IEEE Труды по акустике, речи и обработки сигналов.
2. Датиев М.К., Кулай А.Ю., Датиев К.М. (2013) Новый метод сглаживания вероятностей. Труды молодых ученых. ВНЦ РАН, Владикавказ.
3. Розенфельд Р. (1996) Два десятилетия статистического языкового моделирования. Куда нам идти? Сборник трудов Университета Карнеги Меллон, Питсбург, США.
4. Слеатор Д., Тамперлей Д. (1991) Разбор английского языка при помощи грамматики связей. Технический отчет CMU-CS-91-196, Университет Карнеги Меллон, Питсбург, США.
5. Чен С. (1996) Построение вероятностных моделей для естественного языка. Гарвардский университет.

STATISTICAL LANGUAGE MODELS

M.K. Datiev
graduate student
Vladikavkaz

Moscow Institute of Radio Engineering, Electronics and Automation

A.Y. Kulay
senior researcher
Vladikavkaz

Moscow Institute of Radio Engineering, Electronics and Automation

K.M. Datiev
Cand. Sci. (Engineering), professor
datiev_skgmi@mail.ru
Vladikavkaz

North-Caucasian Institute of
Mining and Metallurgy

Abstract. Modern statistical language models are considered in the article. The applicable criteria of models` efficiency are defined. The following statistical language models are described: n-gramm models, decision tree models, linguistically motivated models.

Keywords: statistical language models, n-gramm models, perplexity.

References

1. Bahl L.R., Brown P.F., de Souza P.V., Mercer R.L. (1989) Statisticheskaia iazy`kovaia model` dlia raspoznavaniia rechi, osnovannaia na derev`iakh [A tree-based statistical language model for natural language speech recognition] IEEE Transactions on Acoustics, Speech and Signal Processing.
2. Chen S. (1996) Postroenie veroiatnostny`kh modelei` dlia estestvennogo iazy`ka [Building Probabilistic Models for Natural Language]. Harvard university.
3. Datiev M.K., Kulay A.Y., Datiev K.M. (2013) Novy`i` metod sglazhivaniia veroiatnostei` [The new method in probability smoothing]. Trudi molodih uchenih. Vladikavkaz.
4. Rosenfeld R. (1996) Dva desiatiletiia statisticheskogo iazy`kovogo modelirovaniia. Kuda nam idti? [Two decades of statistical language modeling: where do we go from here?]. Carnegie Mellon University, Pittsburgh, USA.
5. Sleator D., Temperley D. (1991) Razbor anglii`skogo iazy`ka pri pomoshchi grammatiki sviazei` [Parsing English with a link grammar. Technical Report CMU-CS-91-196]. Carnegie Mellon University, Pittsburgh, USA.

УДК 517.9

ОБ ОПЕРАТОРАХ И УРАВНЕНИЯХ ТИПА РОМАНОВСКОГО С ЧАСТНЫМИ ИНТЕГРАЛАМИ

Ирина Адольфовна Елецких
д.ф.-м.н., доцент
yeletskikh.irina@yandex.ru
г. Елец

Елецкий государственный
университет им. И.А. Бунина

Аннотация. В статье приводится задача теории марковских цепей, поставленная в 1932 году известным советским математиком В.И. Романовским, вводится определение операторов типа Романовского и приводится их классификация. Исследуются линейные операторы типа Романовского с частными интегралами в пространстве непрерывных функций. Свойства таких операторов лежат в основе исследования разрешимости соответствующих уравнений типа Романовского и могут быть использованы при исследовании интегральных уравнений некоторых прикладных задач. Основные результаты получены с применением общей и спектральной теории линейных операторов, а также методов теории интегральных уравнений. В исследовании изучены различные классы таких операторов (с непрерывными, вырожденными, с непрерывными в целом и интегрально ограниченными ядрами) и их пространства. Критерии фредгольмовости и обратимости операторов типа Романовского с перечисленными выше типами ядер применены к изучению условий разрешимости соответствующих уравнений. Изучены композиции операторов и выделен класс операторов, композиции которых