Статистические распределения слов в русскоязычной текстовой коллекции

Statistical distributions of words in a collection of Russian texts

Баглей С. Г. (baglei@galaktika.ru),
Антонов А. В. (alexa@galaktika.ru),
Мешков В. С. (meshkov@galaktika.ru),
Суханов А. В. (sukhanov@galaktika.ru)

Корпорация «Галактика», Москва

Изучение статистических свойств текстов является предметом большого количества работ в области прикладной математики и лингвистики. Подобно многим предыдущим исследованиям, в нашей работе мы придерживались рамок допущения о порождении текста, основанном на случайном процессе Бернулли. Развивая направление, мы исследовали статистические распределения отдельных слов в документах русскоязычной новостной коллекции текстов.

В статье описаны виды распределений слов, относящихся к различным частотным диапазонам. Для наиболее интересующих нас частотных уровней слов проведена аппроксимация графиков распределений, получены коэффициенты функций распределений и величины стандартных отклонений.

Мы рассматриваем полученные статистические данные как основу, на которую можно опираться для получения более реальной оценки вероятности появления некоторого слова в произвольном тексте русского языка. Данная оценка может использоваться для выявления адекватности соответствия слова некоторой текстовой коллекции.

1. Введение

Статистические и вероятностные распределения слов представляют достаточный интерес и являются предметом многих исследований в различных областях: криптографии, статистической теории игр, лингвистике, молекулярной биологии. Несмотря на довольно существенные различия между данными областями, общей для объекта исследования является модель порождения исследуемого текста. В зависимости от области знаний таким текстом может являться естественно-языковой текст, последовательность кодовых слов в цепочке ДНК или передаваемых криптографических данных.

В существующих работах, в основном, описываются два подхода к статистическому представлению текста: один из них основан на Марковском процессе порождения, другой — на модели Бернулли. Использование той или иной модели обусловлено различной природой рассматриваемых предметных областей. В каждом из случаев применения модели формирование данных подчиняется различным принципам и может соответствовать одному из вероятностных подходов в большей или меньшей мере.

В нашем случае предметом рассмотрения был большой набор текстов на русском языке. Каждый из текстов представлял собой реальное новостное сообщение. Мы выбрали для использования модель порождения текста Бернулли, в рамках которой вероятность отдельного испытания (порождения слова) не зависит от результатов остальных испытаний (порождения остальных слов в тексте). В отличие от данной модели, в Марковской цепи порядка N вероятность отдельного испытания зависит от N предшествующих событий.

Для обработки текстов на естественном языке модель Бернулли является более подходящей с точки зрения удобства представления об их формировании. Данная модель лежит в основе многих известных вероятностных алгоритмов обработки текста, таких, как, например, байесовский классификатор текстов [McCallum, 1998.]. При этом, в областях,

не связанных с естественно-языковой обработкой, существует достаточное количество методов, работа которых основана на Марковской модели порождения текста [Schbath, 2000].

Отличие нашего подхода от большинства исследований, опирающихся на модель Бернулли, состоит в том, что мы рассматриваем в качестве некоторого конечного текста не всю текстовую коллекцию, а отдельный документ, являющийся частью набора. Иначе говоря, коллекция в этой модели представляет собой не единый «мешок слов», а объединение нескольких «мешочков». Преимуществом использования данной модели является возможность исследования распределений слов и их частотных характеристик во всем наборе новостных текстов, с учетом элементов набора — отдельных документов.

В своей работе мы попытались получить ответы на некоторые вопросы — например, о том, какие виды распределений наиболее характерны для слов, составляющих большое множество реальных текстов на русском языке. Нам представляется возможным определение достоверности соответствия некоторого подмножества текстов всей текстовой коллекции, основываясь на сравнении соответствия распределения слов в данном подмножестве распределениям этих же слов в коллекции.

Одна из проблем в области корпусной лингвистики, обсуждение которой натолкнуло нас на проведение исследования, и решение которой не удалось определить на «круглом столе» конференции «Диалог'2008», была связана с определением «выбросов» в текстовых корпусах и формулировалась следующим образом. Какая выборка текстов представляется более адекватной в качестве контекста употребления некоторого слова: та, где некоторый термин встречается по 3 раза в 7 текстах, или та, где он встречается по 7 раз в 3 текстах? Данная работа представляется первым этапом для создания инструмента, который может помочь дать ответ на этот вопрос.

2. Распределения слов в текстах

В известных работах в области теории информации [Bayesa-Yates, Ribeiro-Neto, 1999] показано, что частотное распределение слов в большом множестве текстов имеет степенной вид, K/j^{Θ}. Закон Ципфа [Zipf, 1949], описывающий соотношение частоты встречаемости слова и его ранга, в свою очередь, является частным случаем степенного закона распределения, при котором значение Θ близко к единице. В модели большого текстового массива, которую описывает закон Ципфа все множество слов, входящих в коллекцию, рассматривается, как единое целое.

В рамках данного представления проведены исследования [Reinert, Schbath, Waterman, 2000], [Rennie, 2005] о распределениях слов в последовательностях большой размерности n. Было установлено, что распределение частоты встречаемости $N(w)$ некоторого слова w в общем случае подчиняется Гауссову (нормальному) закону распределения. Вместе с тем, оказалось, что такое распределение не наблюдается для тех слов, математическое ожидание появления которых $(n-l+1)\mu(w)$ очень мало. Другими словами, нормальное распределение не наблюдается для редких слов w. Было обнаружено, что более подходящим распределением для таких слов является распределение Пуассона. В качестве объяснения подобного различия в виде распределений можно принимать свойство независимых переменных Бернулли, сумма которых, в зависимости от асимптотического поведения предполагаемого значения, может быть аппроксимирована и в виде нормального распределения, и распределения Пуассона.

Использованная в вышеуказанных работах модель «мешка слов» позволила выявить характеристики распределений, важные для исследований. При этом можно отметить, что за рамками исследований остались характеристики встречаемости слов от текста к тексту внутри коллекций. С точки зрения информационного массива, полностью соответствующего модели Бернулли, данная необходимость, возможно, и отсутствует. Вместе с тем, известно, что естественно-языковые тексты соответствуют модели Бернулли с учетом некоторого приближения. Как правило, слова в текстах на естественном языке распределены неравномерно и исследование подобных «неравномерностей» представляет собой отдельную задачу для изучения.

Распределения слов в отдельных текстах, составляющих текстовые коллекции, также исследовались ранее, например, в следующих работах: [Gotoh, Renals, 2003], [Blake, 2006]. Полученные данные аппроксимировались авторами в различных пространствах, среди прочих, и в пространстве Ципфа «ранг-документ». Было установлено, что у часто используемых общеупотребительных слов распределение частот их встречаемости имеет биномиальный вид. При этом, распределения по документам редких слов, обладающих при этом достаточной статистической базой для построения таких распределений, имеют вид распределений Пуассона, так же, как и при использовании модели единого «мешка слов». Статистическая база для исследований представляла собой в первом случае подборку расшифровок звуковых новостных сообщений объемом 2583 документа, во втором — три словарных корпуса на английском языке объемом, соответственно, 1400, 100 000 и 1 000 000 статей.

Опираясь на результаты, полученные в данных работах, мы исследовали характеристики распределений на большом массиве текстов русского языка. Для нас представляли интерес как задача выявления распределений в применении к текстовому масси-

ву на русском языке, так и анализ распределений на большом массиве данных. Кроме того, нашей целью было получение значений аппроксимирующих функций для проведения дальнейшего исследования в рамках поставленной задачи сравнительного анализа распределений слов в различных текстовых массивах.

3. Результаты экспериментов

Для проведения экспериментов была взята база, содержащая коллекцию новостных сообщений на русском языке за 14-летний период, с 1995 по 2008 год. Основные характеристики базы:

- объем базы — 14,5 миллионов документов;
- средняя длина документа в базе — 503 слова;
- общее количество словомест — 7,3 миллиарда;
- объем словаря базы — 15,6 миллиона слов;
- количество источников-СМИ — около 2 тысяч.

Из всего словаря базы мы отобрали список слов, количество словомест которых превысило 500 вхождений. Данное ограничение было использовано для того, чтобы статистических данных употребления слова было достаточно для построения соответствующего распределения. Отобранные 122,5 тысячи слов были упорядочены по частоте их встречаемости в коллекции.

Далее для каждого слова был сформирован вектор количеств документов, в которых это слово встречается 1, 2, 3, …, 255 и свыше раз. Каждому классу документов в векторе был присвоен свой вес — коэффициент, равный количеству вхождений в него заданного слова. Все случаи, при которых слово встречалось в документах 255 раз и более, были объединены в один общий класс документов, ввиду его априорно достаточно низкого веса. При подсчете количества словоупотреблений мы учитывали все различные морфологические формы, в которых может употребляться некоторое слово.

В результате присвоения весов по количеству вхождений, мы получили для каждого слова еще один взвешенный вектор распределения слова по документам. После получения всех значений была проведена нормировка обоих сформированных векторов.

На рис. 1 в качестве примера приведены распределения значений, полученных для взвешенных векторов слов «прямота», «значение», «быть». Слова были выбраны из различных частотных диапазонов употребления в коллекции: количество словоупотреблений для данных слов отличается от одного к другому на 1,5–2 порядка.

Анализ полученных результатов показал, что распределения по документам группы наиболее частотных слов в коллекции графически соответствуют

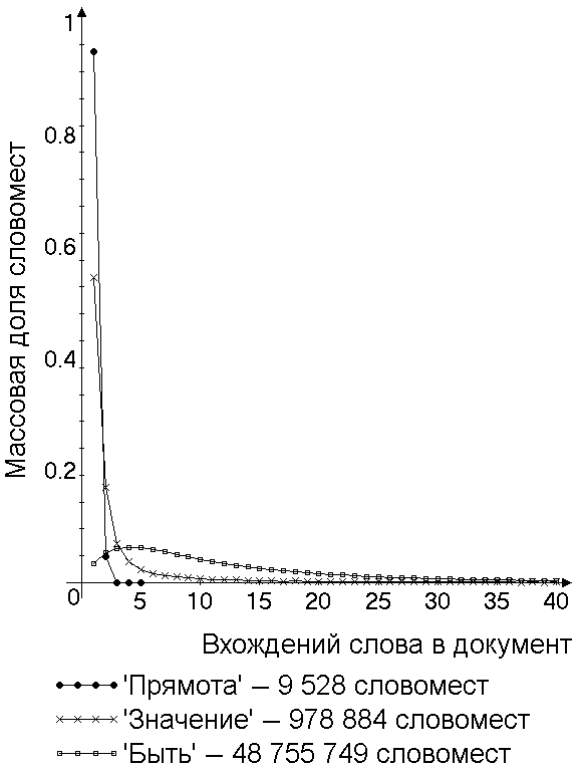

Рис. 1. Распределения взвешенных векторов частот некоторых слов

виду распределения Пуассона с *параметром* λ, значение которого находится в диапазоне 3,5…6,5. При этом усредненное значение количества появлений таких слов в некотором отдельно взятом документе базы также соответствует параметру распределения Пуассона λ, который можно представить в виде математического ожидания количества появлений некоторого слова в документе [Вентцель, 2001]

Распределения прочих слов, не относящихся к наиболее частотным, интересовали нас главным образом на начальном участке координаты вхождений слова в документ. Поскольку на данном участке содержится подавляющая часть значений всего распределения, точность аппроксимации именно в этом диапазоне представлялась наиболее важной с точки зрения практической применимости результатов.

Нами было замечено, что для не самых частотных слов функция распределения Пуассона на начальном участке не обладает достаточной точностью и имеет существенное отклонение в некоторых диапазонах значений. Мы попытались подобрать более подходящую аппроксимирующую функцию, для чего видоизменили график значений распределения частот. Вектор значений частот слов был преобразован к накапливающемуся виду: каждое следующее значение функции стало включать в себя сумму всех предыдущих ее значений.

Далее, чтобы проанализировать влияние частоты словоупотребления на изменение параметров функции, мы: — упорядочили все выбранные слова базы, основываясь на частоте их появления в кол-

лекции; — выделили ряд произвольных интервалов в полученном списке, принимая в качестве критериев выделения следующие условия. Во-первых, диапазон должен иметь статистическую представительность, достаточную для построения распределения. Во-вторых, частоты употребления слов, находящихся на нижней и верхней границах диапазона, не должны существенно отличаться. Допустимый порог для второго из условий составил около 5%. Учитывая данные условия отбора, интересующие нас диапазоны частот соответствовали уровню 10 000, 100 000, 150 000, 500 000, 1 000 000 вхождений слов; — для каждого из интервалов в упорядоченном списке выбранных слов базы была отобрана достаточно объемная выборка слов — по одной тысяче представителей, находящихся в вышеуказанных частотных диапазонах. Чтобы обеспечить равномерность распределения и сгладить погрешности при возможных выбросах, слова выбирались подряд, только с учетом их ранга в упорядоченном списке. Чтобы минимизировать влияние на статистические характеристики выбранного набора слов, мы не учитывали разбиение по прочим признакам — частям речи, ролям в предложениях, формам словоизменения.

После проведения отбора слов мы получили усредненные характеристики распределений слов для заданных частотных интервалов. Мы провели сравнение нескольких методик усреднения и остановились на среднем арифметическом. В отличие от среднего гармонического, среднее арифметическое не вносит своей погрешности в точках нулевых усредняемых значений. По сравнению со средним степенным, среднее арифметическое достаточно удобно с точки зрения его вычисления.

Таким образом, для каждого из списков в 1000 слов были найдены средние арифметические значения элементов взвешенных векторов частот. На рис. 2 приведены графики распределений средних значений по всем исследованным нами интервалам. Для сравнения приведен график, полученный аналогичным образом, для верхушки из 50 слов в коллекции, частота которых превысила 12 миллионов словомест.

Далее с помощью регрессионного анализа значений мы аппроксимировали полученные графики, используя функцию степенного вида:

$$y = 1 - \frac{k}{j^{\theta}} \qquad (1)$$

Выбор функции данного вида был сделан, во-первых, исходя из небольших коэффициентов стандартных (среднеквадратических) отклонений, полученных при аппроксимации. Во-вторых, мы стремились к наибольшей точности аппроксимации на начальном участке графика, так как подавляющая доля значений была представлена именно в этом диапазоне. Пример сравнения графиков значений распределения и аппроксимирующей функции приведен на рис.3.

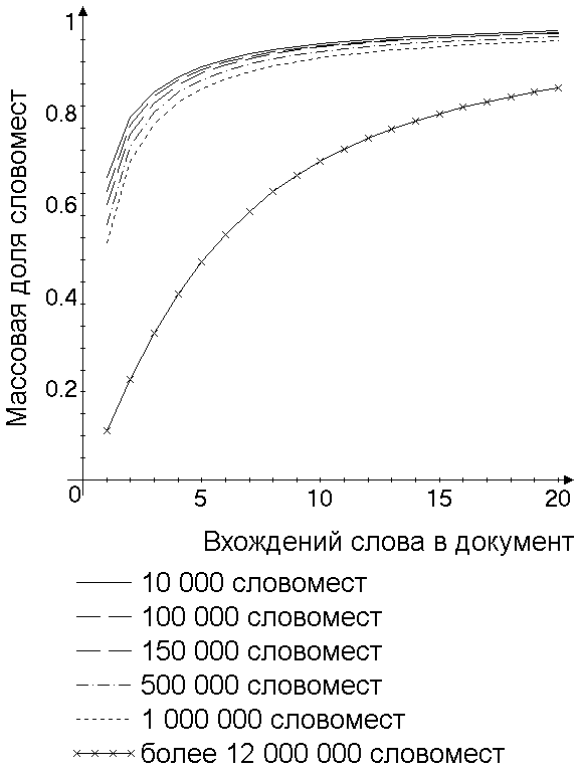

Рис. 2. Распределения накапливающихся взвешенных частот для подмножеств из 1000 слов в некоторых частотных диапазонах

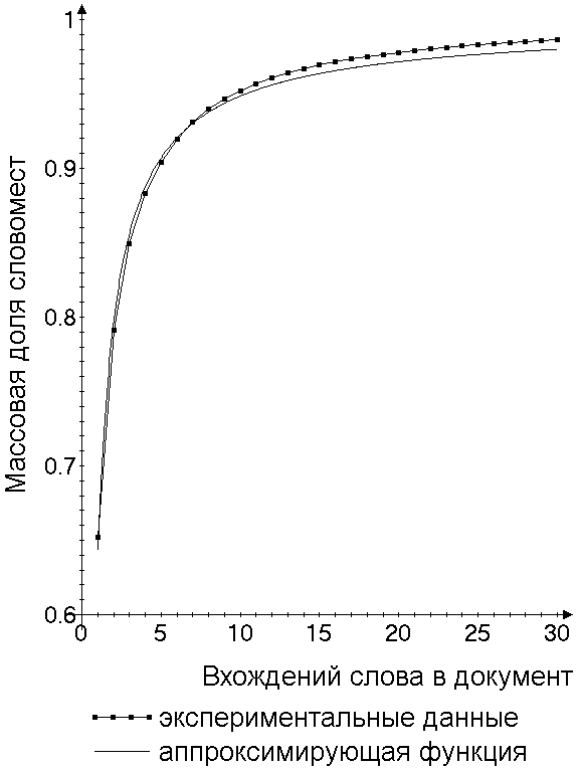

Рис. 3. Аппроксимация распределения накапливающихся взвешенных частот для 1000 слов из диапазона в 50 000 словомест

Аналогичным образом были рассчитаны аппроксимирующие функции для остальных выбранных диапазонов частот. Графики полученных функций приведены на рис. 4.

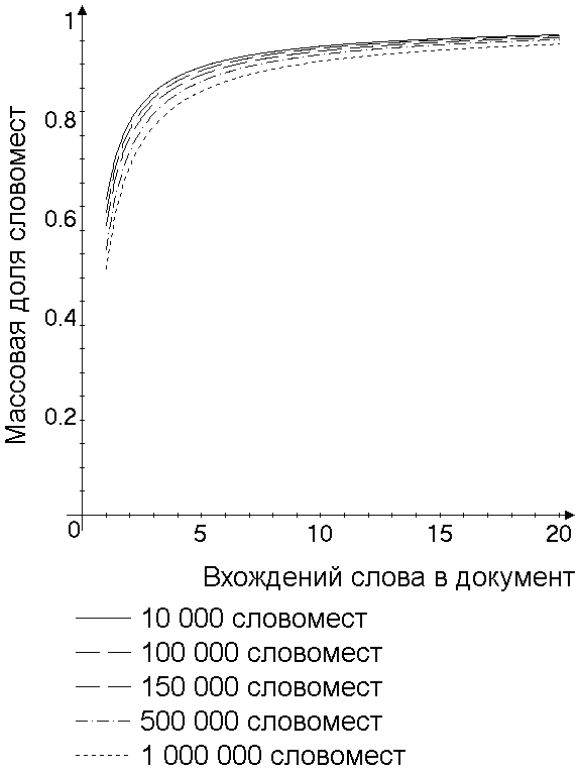

Рис. 4. Аппроксимация распределения накапливающихся взвешенных частот для 1000 слов из заданных диапазонов

В таблице 1 приведены значения коэффициентов функций степенного вида (1), полученных с помощью регрессионного анализа. Коэффициент *s* представляет собой значение стандартного (среднеквадратического) отклонения каждой из функций.

Таблица 1. Значения коэффициентов аппроксимирующих функций

Уровень частот слов	*k*	*Θ*	*s*
10 000	0,370	0,854	0,0109
50 000	0,369	0,857	0,0084
100 000	0,397	0,851	0,0096
150 000	0,424	0,839	0,0106
500 000	0,473	0,838	0,0112
1 000 000	0,512	0,794	0,0101

Очевидно, что коэффициент *k* степенной функции заметно возрастает при увеличении частоты слов в базе. Коэффициент *Θ* при этом убывает. Из полученных данных можно заметить, что для заданных диапазонов частот степенная функция соответствует распределению взвешенных значений частот с приемлемым уровнем отклонений.

4. Выводы

На основе полученных в ходе экспериментов данных можно сделать следующие основные выводы:

- мы получили усредненные параметры распределения некоторых частотных диапазонов слов в большой новостной русскоязычной коллекции текстов;
- на основе анализа списка слов, упорядоченных по частоте употребления в базе, был выделен диапазон частот словоупотреблений, свойственный значимым словам, несущим информацию о различных предметных областях, и имеющий достаточную статистическую базу для построения распределения;
- для некоторых уровней в рамках выделенного диапазона была проведена аппроксимация значений распределения, с помощью которой были найдены коэффициенты функции распределения словомест для заданной частоты словоупотребления в базе.

Полученные результаты не являются окончательными для данной работы. Мы планируем базироваться на них в своих будущих исследованиях, которые можно условно разделить на два этапа.

На первом этапе мы планируем получить более реальную оценку вероятности появления слова в произвольном новостном тексте на русском языке. На втором этапе — использовать функции распределения, полученные в данной работе в сочетании с оценками вероятностей появления слов для задачи расчета оценки соответствия этих слов некоторой произвольной коллекции текстов. Одной из составляющих перечисленных этапов работы может являться исследование влияния грамматических частей речи, а также, возможно, некоторых других признаков слов, на характеристики распределений в текстах.

На наш взгляд, основной целью подобного исследования может являться создание инструмента для получения сравнительных характеристик распределений слов в текстовых массивах. Полученная информация подобного рода может быть применена для выявления наличия или отсутствия аномалий в статистике употребления отдельно взятых слов или словарных групп на основе анализа их распределений в различающихся коллекциях.

Умение правильно использовать такие данные может стать ключевым с точки зрения решения задачи определения приоритета значимости слов, пример постановки которой был приведен во введении данной работы.

Литература

1. *Blake, C.* A Comparison of Document, Sentence, and Term Event Spaces // Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, 2006. Pages: 601–608.
2. *Régnier, M.* A Unified Approach to Word Occurence Probabilities // Discrete Applied Mathematics, 2000. Volume 104, Issue 1–3, Pages: 259–280.
3. *Reinert, G., Schbath, S., Waterman, M.* Probabilistic and Statistical Properties of Words: An Overview // Journal of Computational Biology, 2000. Volume 7, Number 1/2, Pp. 1–46.
4. *Schbath, S.* An Overview on the Distribution of Word Counts in Markov Chains // Journal of Computational Biology, 2000. Volume 7, Number 1/2, Pp. 193–201.
5. *Rennie J.* A Better Model for Term Frequencies // 2005.
6. *Zipf, G.* Human behaviour and the principle of least effort // An introduction to human ecology, 1949. 1st edn., Addison Wesley.
7. *Вентцель, Е.* Теория вероятностей // 7-е изд. стер. М.: Высшая школа, 2001. Т. 2, с. 106–115.
8. *Baeza-Yates, R., Ribeiro-Neto, B.* Modern Information Retrieval // Addison Wesley, 1999.
9. *Gotoh, Y, Renals, S*. Statistical Language Modelling // Lecture Notes in Computer Science, 2003. Springer, Volume 2705, Pages: 78–105.
10. *Régnier, M., Denise A.* Rare events and conditional events on random strings // Discrete Mathematics and Theoretical Computer Science, 2004. Vol. 6, n°2, Pages: 191–214.
11. *Church, K., Gale, W.* Poisson Mixtures // Journal of Natural Language Engineering, 1995.
12. *McCallum, A., Nigam K.* A Comparison of Event Models for Naive Bayes Text Classification // In AAAI/ICML-98 Workshop on Learning for Text Categorization, pp. 41–48. Technical Report WS-98-05. AAAI Press. 1998.