Ю.Д. Голяк (Минск, БГУ)

ДИСТРИБУТИВНО-ЧАСТОТНЫЙ МЕТОД В АВТОМАТИЧЕСКОЙ КЛАССИФИКАЦИИ ТЕКСТОВ

Одной из самых распространенных задач компьютерной лингвистики все чаще становится автоматическая классификация текстов по заданным категориям. И самым точным и гибким инструментом здесь становятся лингвостатистические методы. В данной работе мы попытаемся рассмотреть метод, опирающийся на взаимное окружение элементов текста и их статистические характеристики, а также некоторые реализации этого метода для решения классификационных задач.

Текст воспринимается компьютером как последовательность определенных элементов. Насколько крупные элементы текста будут рассматриваться (от абзаца к предложению, сочетанию слов, слову и отдельному символу) решается в каждом конкретном случае для определенной поставленной цели. При этом основное внимание уделяется не только определению количественных характеристик элементов (частотности), но и учету взаимной дистрибуции элементов текста, поскольку в этом случае элементы искомого сочетания создают взаимный позиционный контекст друг для друга.

На тестовой базе текстов, собранных нами и размеченных или разнесенных по необходимым категориям вручную или же автоматически (там, где это было возможно), высчитываются частоты для каждого из элементов текста. Далее проводится работа по отсеиванию непоказательных результатов. Затем, где необходимо, по определенным математическим правилам высчитывается «удельный вес» наиболее частотных элементов текста для каждой категории.

Таким образом для каждой категории создается специфический словарь маркеров, наиболее показательных элементов, которые с той или иной вероятностью позволяют отнести текст к этому классу. Вероятность находится в прямой зависимости от показателя «удельного веса», присвоенного каждому элементу. После проведения «обучения» программы на тестовой размеченной базе текстов, ей предлагаются непроанализированные тексты. Для каждого из них ведется суммарный подсчет вероятностей вхождения каждого элемента в текст каждого класса из заданных. Наиболее вероятным считается тот класс, для которого сумма вероятностей всех элементов анализируемого текста окажется наибольшей. Необходимо учитывать, что для определения показательных по частоте элементов текста необходим общий частотный словарь языка, на котором записан анализируемый текст. Таково описание этого метода в общем виде.

В рамках разработок Центра прикладной лингвистики, в которых мы принимаем посильное участие, для создания программных приложений автоматической классификации (ниже приводим описание их работы в качестве иллюстрации реализаций описываемого метода) нами была предпринята попытка создания специфического частотного словаря русского языка.

Для этого в качестве текстовой базы для русского языка был взят архив текстов сайта библиотеки Мошкова [4]. Далее была создана программа, анализирующая каждый текст и генерирующая по нему следующие данные: (1) приведенные к начальной форме слова с подсчитанным количеством вхождений в этот текст и частотой для этого текста, с учетом знаков препинания и без такового, а также аналогичные подсчеты для словоформ, (2) сочетания слов с количеством позиций от 2 до 5 (называемых n-граммами, где n — количество позиций в этом сочетании), где каждое слово приведено к начальной форме и без приведения к таковой, с подсчитанным количеством вхождений в этот текст и частотой для этого текста, с учетом знаков препинания, (4) отдельно слова, которых не оказалось в исходном словаре с подсчитанным количеством вхождений в этот текст и частотой для этого текста — таким образом мы получаем статистику по неологизмам, иностранным словам и словам, записанным с

ошибками, (5) и, наконец, файл, полностью повторяющий структуру исходного текста, но содержаший после каждого слова в фигурных скобках морфологические данные об этом слове в закодированном виде.

Для следующего исходного текста:

«Наравне с землей, водой, воздухом и огнем, — деньги суть пятая стихия, с которой человеку чаще всего приходится считаться»

текст с зашифрованной морфологической информацией выглядит так:

«Наравне {наравне: 18. 17409} с {с: 10} землей {земля: 1. 1025. 2051. 3077}, водой {вода: 1. 1025. 2051. 3077. 4098}, воздухом {воздух: 1. 1025. 2049. 3077. 4098} и {и: 11 | и: 37} огнем {огонь: 1. 1025. 2049. 3077}, – деньги {деньга: 1. 1025. 2051. 3074. 4098 | деньга: 1. 1026. 2051. 3073. 4098 | деньга: 1. 1026. 2051. 3076. 4098} суть {суть: 1. 1025. 2051. 3073. 4098. 18433 | суть: 1. 1025. 2051. 3076. 4098. 18433} пятая {пятый: 16. 1025. 2051. 3073. 5122. 7169 | пятая: 1. 1025. 2051. 3073. 4098} стихия {стихия: 1. 1025. 2051. 3073. 4098}, с {с: 10} которой {который: 64. 1025. 2051. 3074. 5122. 7169 | который: 64. 1025. 2051. 3075. 5122. 7169 | который: 64. 1025. 2051. 3077. 5122. 7169 | который: 64. 1025. 2051. 3078. 5122. 7169 | которая: 20. 1025. 2051. 3074. 11267 | которая: 20. 1025. 2051. 3075. 11267 | которая: 20. 1025. 2051. 3077. 11267 | которая: 20. 1025. 2051. 3078. 11267} человеку {человек: 1. 1025. 2049. 3075. 4097} чаще {частый: 2. 6146. 7169 | чаща: 1. 1025. 2051. 3075. 4098 | чаща: 1. 1025. 2051. 3078. 4098} всего {весь: 23. 1025. 2050. 3074 | весь: 23. 1025. 2049. 3074} приходится {приходиться: 5. 1025. 7170. 8194. 9217. 10241. 11267. 12291} считаться {считаться: 9. 7170. 9218. 14337} ».

Полученные данные могут быть использовали для решения следующих задач.

Эмоциональный анализ текста (англ. Sentiment Analysis) — это вид текстового классификационного анализа, предназначенный для выявления в текстах эмоционально окраски и эмоциональной оценки автора по отношению к объектам, речь о которых идет в тексте. Мы приняли за основу принцип бинарного разделения текстов на категории: «positive» (положительная окраска) и «negative» (отрицательная окраска). Алгоритмов эмоционального анализа существует довольно много. В нашем случае был использовали так называемый метод опорных векторов. Этот метод отличается высокой ступенью физической абстракции: здесь сочетания элементов текста, показательные для отнесения текста к тому или иному классу, рассматриваются как объекты в p-мерном пространстве, где p в нашем случае — количество классифицируемых текстов. Эти объекты характеризуются направленностью, то есть с физической точки зрения могут рассматриваться как векторы. Цель такого алгоритма — найти условную плоскость, разделяющую это пространство на два подпространства (в нашем случае бинарного разделения) так, чтобы объекты-векторы оказались максимально удалены от разделяющей линии этой условной плоскости каждый со своей стороны. Таким образом, тексты, для которых характерны эти сочетания элементов (объекты-векторы), оказываются по разны стороны от разделяющей линии этой условной плоскости, то есть происходит классификация. Для обу-

чения программы в качестве опорной базы текстов были взяты 10000 размеченных по двум заданным категориям сообщений из социальной сети Twitter. Далее при работе с ними был применен так называемый метод кросс-валидации. Суть этого метода заключается в следующем: размеченная база делится условно на десять частей; с некоторых пяти частей разметка снимается и эти пять частей становятся тестовой базой, другие пять частей — обучающей эталонной базой. После обучения на эталонной базе запускается классификация на тестовой, и результаты сравниваются с тем, как тестовая часть была изначально эталонно размечена. Затем тестовая и эталонная базы комбинируются из некоторых других пяти частей общей базы, и весь цикл повторяется несколько раз — столько, сколько комбинаторно возможно из десяти сегментов исходной базы, то есть 252 раза. Таким образом точность программы была доведена до 96,1%.

Задача *тематической классификации* текстов сходна с задачей эмоционального анализа по сути, но принципиально различается по количеству заданных для классификации классов. Такая классификация — автоматическое отнесение текста к тому или иному семантическому классу — может быть необходима во многих случаях: такую рубрикацию используют новостные сайты для автоматического распределения статей по темам; это необходимо при создании электронных архивов документов, когда традиционно встает задача упорядочения информационного массива, объединения его в группы, называемые категориями, рубриками, тематическими подборками, кластерами, сюжетами и т.д. Реализация здесь дистрибутивно-частотного метода почти не отличается от предыдущего случая: для отнесения текста к тому или иному классу высчитываются вероятности вхождения отдельных частей текста (отдельных слов или их сочетаний, и тогда элементы, формирующие эти сочетания являются взаимной дистрибуцией друг для друга). Однако принимая во внимание тот факт, что классов при такой классификации может быть задано почти неограниченное количество, метод опорных векторов здесь может привести к значительному усложнению вычислений. Поэтому для решения этой задачи лучше всего подошел метод максимальной энтропии, в котором вероятность отнесения текста к одному из заданных классов описывается в общем виде следующей формулой (1):

$$P(c|d,\lambda) \stackrel{def}{=} \frac{\exp\sum_i \lambda_i f_i(c,d)}{\sum_{c' \in C} \exp\sum_i \lambda_i f_i(c',d)}, \quad (1)$$

где λ_i — вес каждого i-того слова, а f_i рассчитывается для каждого i-того слова по формуле (2).

$$f_i(w_i|c) = \frac{f(w_i,c)+1}{f(c)+|V|}, \quad (2)$$

где w — слово, c — класс, $f(w,c)$ — количество вхождений слова в тексты этого класса, $f(c)$ — количество слов в этом классе, $|V|$ — размер словаря, на который опирается алгоритм.

Для тестовой классификации были заданы 26 тематических категорий (Авто, Бизнес, Дизайн, Здоровье, Игры, Интернет, История, Кино и ТВ, Книги, Компьютеры, Космос, Красота, Мода, Музыка, Наука, Обучение, Отдых, От-

ношения, Политика, Природа, Путешествия, Спорт, Религии, Экономика, Техника, Электроника), по которым вручную нами были разнесены 6000 текстов. В эту базу вошли тексты художественного стиля и новостные интернет-статьи. Точность работы программы составляет 87-88%, что, безусловно, говорит о необходимости ее усовершенствования, однако для решения некоторых задач она уже пригодна.

Кроме того дистрибутивно-частотный метод может быть использован при *автоматическом определении языка*. Одним из вариантов анализа текста является поиск в нем самых характерных, частотных сочетаний символов, позволяющих отнести текст к тому или иному языку. Символы в таких сочетаниях можно рассматривать как взаимную дистрибуцию по отношению друг к другу, а их частотность является показателем для определения языка. Безусловно, такой метод подразумевает наличие опорного списка частотных символьных сочетаний с их частотами для каждого поддерживаемого программой языка. В нашем случае это 9 языков (африкаанс, немецкий, английский, сербский (лат.), сербский (кирилл.), французский, итальянский, польский, белорусский, русский), среди которых один — сербский — регулярно встречается и в кириллическом, и в латиническом варианте. На основании базы текстов на каждом языке были созданы списки самых частых сочетаний символов от 2 до 5 позиций включительно. Точность работы такого алгоритма составляет 98.3%.

ЛИТЕРАТУРА

1. Андреев, Н. Статистико-комбинаторные методы в теоретическом и прикладном языковедении / АН СССР. Ин-т языкознания. — Л., 1967.

2. Плотников, Б. Дистрибутивно-статистический анализ лексических значений. — Минск, 1979.

3. Daniel Jurafsky, James H. Martin Speech and language processing: an introduction to natural language processing, computational linguistics and speech recognition. [Электронный ресурс]. — Режим доступа: http://www.cs.colorado.edu/~martin/slp.html. – Дата доступа: 02.03.2013.

4. Электронная библиотека. [Электронный ресурс]. – Режим доступа: http://www.lib.ru/. — Дата доступа: 02.03.2013.