УДК 004.852

АНАЛИЗ ТЕКСТОВ ДЛЯ ФОРМИРОВАНИЯ ЭЛЕКТРОННЫХ ЧАСТОТНЫХ СЛОВАРЕЙ

Середин А. И.
научный руководитель д-р техн. наук Ковалев И. В.
Сибирский Федеральный Университет

Формирование информационно-терминологического базиса (ИТБ) для определенной области знаний происходит на основе частотных словарей, относящихся к этой области. Поэтому важным этапом перед формированием ИТБ является подбор определенного количества языкового материала (текстов), посвященного рассматриваемой области знаний и составление на его основе частотных словарей. Для решения этой задачи широко применяются различные статистические методы.

Статистические методы все шире и глубже проникают в самые различные области научной деятельности, в том числе и в такие, традиционно считавшиеся “неточными”, как языкознание. Совокупность этих методов, используемых в науке о языке, позволяет определить, какие языковые явления встречаются в речи или тексте чаще, а какие реже. Наиболее интенсивно статистическими методами изучается словарный состав языка. Сведения об употребительной лексике дают частотные словари. Частотный словарь регистрирует слова, словоформы или словосочетания, которые встретились в исследованном для его составления тексте (выборке). Для этих единиц (т.е. слов, словоформ или словосочетаний) в словаре указываются их частоты, т.е. числа, показывающие, сколько раз каждая словарная единица встретилась в данном тексте.

Составление частотного словаря требует значительных затрат времени и знакомства со статистической методикой наблюдений. В докомпьютерную эпоху составление частотного словаря происходило вручную. Например, создатели частотного англо-русского словаря-минимума газетной лексики придерживались следующей методики. Словарь составлялся на основе лингвостатистического анализа языка газет и журналов Великобритании и США. Были отобраны тексты общей длиной 200000 словоупотреблений (под словоупотреблением в лингвостатистике принято понимать слово как единицу длины текста) из разных газет и журналов. Из этих текстов вручную выписывались слова и словосочетания с различной степенью устойчивости, и анализировалась частота их употребления. В наше же время создание частотных словарей можно автоматизировать с помощью компьютера. Таким образом, значительно сокращается время, необходимое на составление частотного словаря, а также повышается его точность.

Текст по своей природе обладает статистической структурой. Сущность ее состоит, с одной стороны, в том, что все слова и выражения, составляющие словарь текстов для данной области знаний, а также грамматические формы и синтаксические конструкции, обладают определенной вероятностью появления в текстах этой области знаний.

С другой стороны, если разбить рассматриваемый текст на малые порции, то одна часть лингвистических единиц будет давать в этих текстах примерно одни и те же частоты, таким образом, показывая, устойчивость и равномерность употребления. Другая группа лингвистических единиц дает неустойчивое и неравномерно употребление в отдельных порциях текста. Первую группу обычно составляют служебные слова и общеупотребительные словосочетания. Вторую группу образуют

чаще всего слова и словосочетания, непосредственно связанные с содержанием текста данной области знаний (эти слова и словосочетания часто называют ключевыми).

Структура текстов разных областей знаний неодинакова: в них заметно отличаются вероятности и распределения ключевых слов и словосочетаний, статистика же служебных слов и некоторых общеупотребительных слов и словосочетаний остается практически неизменной.

Если исследовать текст достаточно большого объема, можно обнаружить закономерности функционирования лексики данного языка и получить представление о ее количественной структуре. При таком анализе выявляются, например, две важнейшие лингвостатистические закономерности.

Первая закономерность состоит в том, что, в любом тексте, каким большим бы он не был, используется лишь незначительная часть словарного богатства языка.

Например, специальные научно-технические и публицистические тексты весьма отличаются по объему словаря. Анализ текстов на английском, румынском и молдавском языках показал, что словарь текстов публицистического характера примерно в 2,5 раза больше, чем словарь специальных текстов.

Эти цифры свидетельствуют о том, что в разных сферах речевого общения используются и разные количества слов.

Вторая лингвостатистическая закономерность состоит в том, что даже и ограниченная часть лексики языка используется в речи (тексте) неравномерно. Одни слова употребляются чаще, другие реже, причем большая часть всего текста приходится на незначительное количество самых частых слов. Например, при записи и анализе телефонных разговоров были получены следующие результаты: 737 самых частых слов занимают свыше 95% всех словоупотреблений.

Как уже отмечалось, в частотном словаре указывается количество случаев употребления слова в тех текстах, которые были проанализированы для составления словаря. Частотные словари различаются в зависимости от принципа размещения материала. Слова или словосочетания могут быть расположены по алфавиту – как в обычном словаре, с проставлением рядом со словом его частоты. Также слова и словосочетания могут быть расположены по убыванию частот, начиная от самого употребительного слова. Первый вариант частотного словаря предназначен для обучаемого, второй – обучающему. Обучаемый может также работать со вторым вариантом словаря при самостоятельном изучении иностранного языка, например, при заучивании слов и словосочетаний порциями в зависимости от их частоты или при проверке владения словарными единицами, начиная с самых частых.

При автоматизации общего статистического анализа могут быть выделены следующие этапы:

- определение статистических элементов (слово, фраза, предложение);
- определение абсолютной частоты элементов по единичной выборочной пробе и общей выборочной пробе;
- расчет относительной частоты и вероятности для основной совокупности терминов определенной области знаний;
- проверка достоверности полученных частотных характеристик путем вычисления стандартных отклонений и относительной ошибки;
- формализация результатов в виде списков, таблиц или графиков;
- интерпретация и обобщение результатов, вплоть до формулирования закономерностей.

Так как практически невозможно охватить всю общность предметно-языковой коммуникации даже только для одного языка и одной области, предметно-языковая статистика должна опираться на наиболее репрезентативные выборочные пробы, т.е. на

письменные или устные предметно-типичные тексты. Каждый языково-статистический анализ начинается с выбора и подготовки соответствующей текстовой базы. При специфических постановках задач в рамках прикладного языкознания, например, при определении словарного запаса для заучивания на занятии по иностранному языку или при составлении вокабуляра для внутрипроизводственной документации, объем текстовой базы может быть сильно ограничен.

Необходимо также обращать внимание на вид текстов. Особенно пригодны для определения научно-технического основного словарного запаса учебники высшей и профессиональной школы обзорного характера. Они гарантируют систематический, пропорциональный и полный охват материала и необходимые языковые средства для его изложения, кроме того, они в меньшей степени подвержены влиянию со стороны индивидуального языкового употребления отдельных представителей профессии. Дальнейшее формирование текстовой базы основывается на использовании новых журналов не специального характера.

Первым результатом статистической обработки текста является абсолютная частотность. Она показывает, как часто возникает соответствующее явление в исследуемом тексте. Однако, она имеет малую ценность для дальнейших исследований при практическом использовании результатов или вообще для обобщенных высказываний, так как она напрямую зависит от объема выбранного текста. Она служит исключительно как исходная величина, например, для расчета относительной частотности.

Относительная частотность – процентная величина, которая выражает долю языковой единицы в целом тексте. Она получается из деления абсолютной частотности на длину выборочной пробы (1).

$$f_r = \frac{f_a}{N}, \tag{1}$$

где f_r – относительная частотность, f_a – абсолютная частотность, N – длина выборочной пробы.

Например, для слова с абсолютной частотностью 173 в одной выборочной пробе из N=60000 языковых единиц, относительная частотность будет вычисляться как 173/60000=0,00288.

Другими словами, относительная частотность явления – отношение числа его действительного возникновения к числу его теоретически возможного появления. Если выборка по величине репрезентативна для предметного языка, тогда можно приравнять относительную частотность к вероятности языкового явления. Тогда она дает основание для выводов о статистической структуре соответствующего субъязыка или о важности отдельных элементов для организации текста.

Особенно важным шагом при статистическом анализе языка является контроль достоверности определяемых данных. Для этого в распоряжении имеются различные способы контроля. В предметно-языковой статистике учитываются, прежде всего, стандартные отклонения (погрешности), относительная ошибка и конфиденциальные границы.

Стандартная погрешность (средняя квадратная погрешность) – мера изменчивости средней частотности языкового явления в частичных выборочных пробах. Она рассчитывается по формуле:

$$S = \sqrt{\frac{SAQ}{n-1}}, \tag{2}$$

где S – стандартная погрешность, SAQ – сумма квадратов погрешностей, n – число контрольных проб.

Относительная ошибка вычисляется, прежде всего, для определенных лексических единиц в частотных словарях, чтобы определить достоверность этих словарей. Общепринятая формула определения относительной ошибки:

$$\delta = \frac{Zp}{\sqrt{nf}}, \tag{3}$$

где δ - относительная ошибка, Zp – коэффициент для данного уровня доверия p, n – объем выборки (выборочной пробы), f – относительная частотность.

Расчет интервала доверия – это уточненный вариант расчета относительной ошибки, с которой определяется нижняя и верхняя граница (p_1 и p_2) колебаний и средняя частотность. Существуют разные способы расчета интервала доверия, например:

$$p_1 = \frac{fN + \frac{1}{2}Zp^2 - Zp\sqrt{f(1-f)N + \frac{1}{4}Zp^2}}{n + Zp^2}, \tag{4}$$

$$p_2 = \frac{fN + \frac{1}{2}Zp^2 + Zp\sqrt{f(1-f)n + \frac{1}{4}Zp^2}}{n + Zp^2}. \tag{5}$$

При отображении результатов исследований обычно используются различные списки, таблицы, графики и т.п. С помощью кругового изображения и ленточных диаграмм изображаются части в процентных величинах. Для графического изображения количественных признаков, таких как длина слова или предложения, пригодны гистограммы и цепь многоугольников. Графики-кривые с более или менее типичным течением по качественным и количественным признакам превышают это простое сочетание частотностей. Они позволяют распознавать функциональные связи между признаками и их частотностью, и частотность языковых явлений сама может стать признаком того, что характеризуется другими данными.

Таким образом, частотные словари позволяют учитывать статистические свойства текстов, что в свою очередь приводит к построению на их основе более качественных и полных информационно-терминологических базисов. Но успешное составление частотных словарей очень сильно зависит от этапа подбора репрезентативной выборки текстов по определенной области знаний. Этот процесс относительно плохо автоматизируется. Зато, уровень развития современных информационных технологий позволяет выполнять составление частотных словарей и последующее формирование ИТБ на их основе, практически полностью автоматизировано. Что позволяет существенно сократить затраты времени и человеческих ресурсов на их составление. А это, в свою очередь, ведет к ускорению и удешевлению процесса разработки систем компьютерного обучения языкам.