УДК 624.012:801.52

А.Т.АБДЫШОВА
E-mail: ksucta@elcat.kg.

СПОСОБЫ И ФОРМЫ МОДЕЛИРОВАНИЯ ТЕХНИЧЕСКОГО ТЕКСТА

Бул макалада кептин мүмкүндүк-статистикалык сүрөттөлүшүнүн колдонушу менен тилдин машиналык базалык түзүлүшү, текстин күтүлбөгөн процесс катары каралышы, азыркы аппаратты изилдөө үчүн колдонулушуна мүмкүндүк берет, ошондой эле тексттин автоматтык иштетилишинин түзүлүшү үчүн керектүү тигил же бул программаларынын ушул изилдөөлөрдүн негизинде объективдүү лексиканы жана морфологияны тандоону уюштурат.

В данной статье рассматривается построение машинного базового языка с применением вероятностно-статистического описания речи, где текст рассматривается как случайный процесс, что делает возможным применение для исследования построенной модели математического аппарата, а также позволит организовать на основе этих исследований объективный отбор лексики и морфологии, необходимой для построения тех или иных программ автоматической переработки текста.

In given article construction of machine base language with application of the is likelihood-statistical description of speech is considered where the text is considered *as casual process that does possible application for research of the constructed model a mathematical apparatus, and also to organise on the basis of these researches objective selection of lexicon and the morphology necessary for construction of those or other programs of automatic processing of the text.*

Общение человека с ПК – в том числе и с совершенными машинами большой мощности – может осуществляться лишь при условии, что в память компьютера будет введена некоторая модель, представляющая собой сокращенное описание естественного языка. Эту модель, включающую наиболее существенные с точки зрения текстообразования его элементы и связи, мы будем называть машинным базовым языком (БЯ) /1/.

Построению машинного БЯ предшествует вероятностно-статистическое описание речи, выявляющее основные элементы и текстообразующие свойства конкретного языка. Это описание осуществляется с помощью последовательного использования различных вероятностно-статистических и информационных моделей все возрастающей сложности. Основная идея вероятностно-статистического моделирования состоит в том, что текст рассматривается как случайный процесс, а единицы текста (буквы, буквосочетания, морфемы, словоупотребления, словосочетания, грамматические схемы) – как случайные события /2/.

Каждый случайный процесс является результатом действия некоторой производящей системы. Для текста такой производящей системой являются система, норма и узус языка или

подъязыка. Появление лингвистической единицы, являющейся случайным событием, предполагает некоторый комплекс условий, при выполнении которых осуществляется или не осуществляется данное лингвистическое событие. В этот комплекс входит описываемая ситуация, жанровые особенности текста и личность автора, окружающий контекст и т.д. При моделировании текста комплекс условий определяется в зависимости от поставленной задачи. Это значит, что мы заранее отказываемся от описаний менее важных для нас сторон текста с тем, чтобы получить возможность моделировать те его свойства, которые особенно важны с точки зрения доставленной задачи.

Наиболее простой вероятностно-статистической моделью текста, использующейся при обучении вычислительных машин, является частотный список (словарь) словоформ и слов .

Частотные словари (ЧС) слов и словоформ строятся исходя из следующих упрощающих допущений /3/.

1. Моделирующий текст и текст, выступающий в роли натурного объекта, рассматриваются как реализация стационарного случайного процесса. Предполагается, что производящие текст система, норма и узус, а также сопутствующий им комплекс условий остаются неизменными, в связи с чем распределение вероятностей в тексте для всех рассматриваемых лексических единиц также остается неизменным. В действительности каждый реальный текст является нестационарным процессом. Распределение вероятностей употребления многих слов и словоформ зависит от темы текста, времени его написания (эта зависимость имеет место даже в том случае, когда речь идет о коротком хронологическом периоде), от возраста, образования и стилевых вкусов автора.

2. Моделирующий текст рассматривается как последовательность независимых испытаний, не учитывающих вероятностные связи (лексико-грамматические и стилистические валентности), существующие между словоупотреблениями текста. Хотя оба упрощающих предположения делают текстовую модель мало похожей на реальный текст и приводят к заметным потерям смысловой и синтаксической информации, они дают возможность применить для исследования построенной модели математический аппарат, а также организовать на основе этих исследований объективный отбор лексики и морфологии, необходимой для построения тех или иных программ автоматической переработки текста.

Автоматическая переработка текста, учитывающая актуальные значения составляющих его словоформ, становится возможной при условии, если используемая в ЭВМ лингвистическая модель включает списки типовых контекстов и валентностей.

Словосочетания, реализующие типовые контексты и дистрибуции, чаще всего представляют собой трехсловные (триады) либо двухсловные (диады), четырехсловные (тетрады), пятисловные (пентады) и т.д. цепочки знаменательных и служебных слов. Выборка этих словосочетаний в ЧС может производиться вручную или на ЭВМ с помощью иконической или эвристико-алгоритмической процедур.

Поэтому для получения достоверных статистических данных даже о наиболее частых триадах и тетрадах пришлось бы использовать выборки текста, во много раз превосходящие выборки, применяющиеся при построении ЧС словоформ и слов. Во-вторых, получаемый список содержит десятки и даже сотни тысяч случайных бессмысленных комбинаций, а также не полностью оформленных цепочек, которые не всегда удается однозначно отделить от семантически и грамматически оформленных словосочетаний.

Более экономным является целенаправленный выбор из текста триад и тетрад, опирающихся на заранее заданные опорные слова-ядра, которые находятся в строго фиксированной позиции относительно остальных элементов словосочетания.

Частотные словари составляются на основе экспериментальных выборок, каждая из которых отражает лишь малую часть текстов, входящих в генеральную совокупность. Поэтому в частотные словари не попадает большое число редких слов или словоформ, образующих массив резервных лексических единиц, каждая из которых имеет в данном частотном словаре F=0. Не попадая в составляемый на основе ЧС автоматический словарь, эти нуль-частотные слова и словоформы образуют находящийся за пределами машинного БЯ лексический массив, являющийся источником той частотной непокрываемости текста в 1–3 %, о которой говорилось выше. Эти непознаваемые автоматическим словарем нуль-частотные единицы являются обычно теми высокоинформативными словоформами, которые необходимы для формального распознавания текста. Нуль-частотные слова и словоформы входят также в сложные лексические единицы, которые также служат средством выражением новизны текста.

Каждая область знаний использует несколько десятков, а иногда и сотен тысяч простых и сложных терминов. Так, например, в русском языке при обработке «исчерпывающей представительной коллекции информационных документов по химии, химической и нефтехимической промышленности» выделено до 180 тыс. сложных и простых терминов. Обычно считается, что биология и медицина используют каждая до миллиона сложных и простых терминов, включая греко-латинские образования. Оценивая объем терминологической лексики в отдельном подъязыке, можно воспользоваться опытом построения автоматических двуязычных словарей. Однако полные словники отраслевых словарей в аналитических языках включают десятки тысяч, а в синтетических языках, вероятно, сотни тысяч словоформ. Отсюда следует, что нуль-частотные лексические единицы составляют в каждом из подъязыков значительный массив словоформ. Хотя эти словоформы и покрывают всего 2-3 % контрольного текста, они, с одной стороны, сами выражают новые научно-технические понятия, а с другой – участвуют в формировании сложных терминов, также использующихся для воплощения ремы текста.

Наряду с формированием грубых статистических моделей – частотных списков – в лингвистике текста делались попытки построить два типа вероятностных моделей текста. Авторы моделей первого типа строили их исходя из следующих рассуждений.

Соотношением $m(F, N) = G(N) F^{-(1+a)}$ описывается ряд ситуаций, где $a > 0$, а $G(N)$ – коэффициент, зависящий от объема выборки N, встречающихся в разных отраслях знаний. Этой

закономерности подчиняется отношение между числом людей, получающих определенный доход, и величиной этого дохода (так называемый закон Парето); она описывает отношение между числом родов m (F, N), содержащих ровно F видов, и самой величиной F (так называемый закон Уиллиса-Юла). Применительно к лингвистике эта закономерность связывает число словоформ (или слов) m, встретившихся ровно F раз в выборке N с самой частотой F.

Вторая серия вероятностных моделей текстообразования, наиболее полно изложенная в работах Б.Мандельброта, строилась исходя из более сильных теоретических посылок, учитывавших, с одной стороны, стохастические схемы Маркова, а с другой – идеи шенноновской теории информации. Теоретические построения Б.Мандельброта дают возможность применить для исследования построенной модели математический аппарат, а также организовать на основе этих исследований объективный отбор лексики и морфологии, необходимой для построения тех или иных программ автоматической переработки текста.

Выявление доминантных единиц и единиц заполнения в тексте можно выявить применением вероятностно-статистических методов. Рассмотренные выше статистические и вероятностные модели строились исходя из упрощающего допущения о том, что текст представляет собой реализацию стационарного процесса. В роли генератора этого процесса выступают система языка, его норма и узус, которые, согласно этому допущению, остаются неизменными в рамках заданного синхронного среза. В действительности в порождении текста участвуют не только система, норма и узус языка, но также вечно меняющаяся ситуация. Можно предположить, что те лингвистические элементы, появление которых определяется в основном стационарной системой, нормой и узусом языка, будут иметь в тексте иные статистические характеристики, чем те единицы, появление которых подсказывается нестационарной ситуацией.

Для проверки этого предположения был использован аппарат теории распределений. Идея этой проверки состоит в определении степени схождения эмпирических вариационных рядов отдельных лингвистических единиц с тремя теоретическими моделями:

а) распределением Пуассона (распределением редких лингвистических элементов)

$$p(F) = \frac{\lambda^F}{F!} e^{-\lambda},$$

(1)

где P (F) – вероятность появления данной лингвистической единицы ровно F раз в серии из N единиц, а λ – параметр формулы, оцениваемый с помощью средней частоты исследуемого лингвистического элемента;

б) нормальным распределением

$$P(F) \approx \frac{1}{\sqrt{2\pi Nf(1-f)}} \exp\left[- \frac{(F - Nf)^2}{2Nf(1-f)} \right],$$

(2)

где f – относительная частота (эмпирическая вероятность) интересующей нас лингвистической единицы (остальные величины имеют тот же смысл, что и в распределении Пуассона);

в) логнормальным распределением

$$p(F) \approx \frac{1}{F_\sigma \sqrt{2\pi}} \exp\left[-\frac{(\ln F - p)^2}{2\sigma^2} \right], \qquad (3)$$

где $p = \dfrac{\sum_{i=1}^{N} \ln F_i}{N}$, а $\sigma = \dfrac{\sqrt{\sum_{i=1}^{N} (\ln F_i - \mu)}}{N - 1}$.

Рассмотрены статистические и вероятностные модели обработки текста исходя из упрощающего допущения о том, что текст представляет собой реализацию как стационарного, так и нестационарного процесса для дальнейшего его использования в науке.

СПИСОК ЛИТЕРАТУРЫ

1. Борисевич А.Д., Крисевич В.С.. Частотный словарь словоформ английского подъязыка строительных материалов //Статистика текста. – Минск: Б ГУ, 1969.
2. Букович В.А. Частотный словарь английского подъязыка электронно-вычислительной техники //Статистика текста. – Минск: Б ГУ, 1969.
3. Тарасова Е.С. Частотный словарь английских текстов. //Статистика текста. – Минск: Б ГУ, 1969.
4. Ем Н.В. Частотный список английского подъязыка физической химии //Статистика текста. – Минск: Б ГУ, 1969.
5. Выборка 50 тыс. словоупотреблений. Разных словоформ 5177. Текст обработан вручную.
6. Мелик-Гусеинова Р.С. Частотный словарь английских текстов по физике твердого тела //Статистика текста. – Минск: Б ГУ, 1969.
7. Выборка 100 тыс. словоупотреблений. Разных словоформ-5542. Текст обработан вручную.
8. Кутуев М.Д., Укуев Б.Т. Моделирование недетерминированных факторов.
9. Информационные технологии в строительной механике /М.Д.Кутуев, Б.Т.Укуев, С.Е.Ешпулатов, Р.А.Куканова, А.Г.Шубович.