ОБОЗРЕНИЕ
ПРИКЛАДНОЙ И ПРОМЫШЛЕННОЙ
Том 20 МАТЕМАТИКИ Выпуск 5
2013

**С. Ю. Мельников** (Москва, ООО «Линфо»). **Определение языка текста как задача идентификации вероятностного автомата.**

Определение языка текста — задача отнесения имеющегося текстового фрагмента, представленного в конечном алфавите, тому или иному естественному языку из заданного перечня. Задача является актуальной для ряда приложений, и для ее решения предложены практически эффективные подходы ([1]).

Задача идентификации конечного (вероятностного) автомата по наблюдениям над выходной последовательностью относится к классическим задачам математической кибернетики. Под задачей идентификации мы понимаем задачу проверки гипотезы о том, что неизвестный автомат $A'$, выходная последовательность которого наблюдается, совпадает с эталонным автоматом. При этом неизвестный автомат выбирается из заранее заданного класса ([2], [3]).

Первая задача — априори неформализуема и поэтому все существующие методы ее решения являются эвристическими, и могут подтверждаться только практическими вычислениями на текстовых корпусах. Вторая задача — сугубо формальная, допускающая точное математическое решение.

Имеется определенная близость между статистическими методами, которые используются для решения задачи идентификации языка и методами, предложенными для идентификации конечных автоматов. Подавляющее большинство методов, которые применяются для решения обоих задач, основаны на подсчете $n$-граммных статистик символов и слов в выходном алфавите.

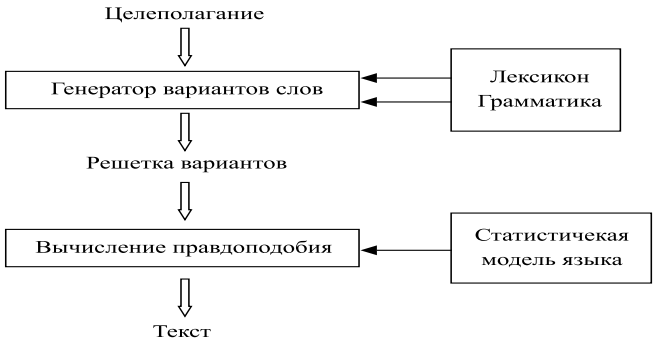


Рис. Типовая схема генератора текста

На сегодняшний день разработан и активно применяется целый ряд схем синтеза текста ([4]). Реализации таких схем используются при составлении читаемых текстов в различных предметных областях (прогнозы погоды, составление описаний музейных экспонатов, описание ситуаций в компьютерных играх, генерация веб-страниц и др.). В ряде систем происходит синхронная генерация текстов на нескольких языках. Наиболее совершенные генераторы используют богатые лингвистические базы,

содержащие грамматические и статистические знания о языке. В докладе рассматриваются структурные схемы таких генераторов. Показано, что их математическими моделями могут служить конечные вероятностные автоматы. Как правило, чем лучше генератор текста, тем сложнее его автоматная модель. Элементы случайности в вероятностной модели связаны, во-первых, со случайностью входных данных, поступающих на вход генератора текста (целеполагание, т. е. что именно должен отражать генерируемый текст), и, во-вторых, с неоднозначностью выбора лингвистических ресурсов, использованных при построении генератора. Отсюда можно сделать вывод, что трудноформализуемая задача идентификации языка естественного текста имеет своим непосредственным аналогом задачу идентификации вероятностного автомата.

## СПИСОК ЛИТЕРАТУРЫ

1. *Максимов А. В.*, *Мельников С. Ю.*, *Чавчавадзе Н. М.* Тенденции развития методов автоматической идентификации языка речевых и текстовых сообщений. — Обозрение прикл. и промышл. матем., 2009, т. 16, в. 2, с. 365–367.
2. *Кудрявцев В. Б.*, *Грунский И. С.*, *Козловский В. А.* Анализ и синтез абстрактных автоматов. — Фундамент. и прикл. матем., 2009, т. 15, в. 4, с. 101–175.
3. *Мельников С. Ю.* Идентификация конечных автоматов на основе метода многогранников. М.–Ижевск: Ин-т компьютерных иссл., 2013, 136 с.
4. *Reiter E.*, *Dale R.* Building Natural Language Generation Systems. Cambridge: Cambridge Univ. Press, 2000, 145 p.