ISSN 1684–8853

ИНФОРМАЦИОННО-УПРАВЛЯЮЩИЕ СИСТЕМЫ

НАУЧНЫЙ ЖУРНАЛ

4(47)/2010

НАУЧНЫЙ ЖУРНАЛ

4(47)/2010

РЕЦЕНЗИРУЕМОЕ ИЗДАНИЕ

ИНФОРМАЦИОННО-УПРАВЛЯЮЩИЕ СИСТЕМЫ

Учредитель
ОАО «Издательство “Политехника”»

Адрес редакции: 190000, Санкт-Петербург,
Б. Морская ул., д. 67, ГУАП, РИЦ
Тел.: (812) 494-70-44
Факс: (812) 494-70-18
E-mail: 80x@mail.ru
Сайт: www.i-us.ru

Журнал зарегистрирован в Министерстве РФ по делам печати, телерадиовещания и средств массовых коммуникаций.
Свидетельство о регистрации ПИ № 77-12412 от 19 апреля 2002 г.

Журнал входит в «Перечень ведущих рецензируемых научных журналов и изданий, в которых должны быть опубликованы основные научные результаты диссертации на соискание ученой степени доктора и кандидата наук».

Журнал распространяется по подписке. Подписку можно оформить через редакцию, а также в любом отделении связи по каталогам:
«Роспечать»: № 48060, № 15385; «Пресса России»: № 42476.

ЛР № 010292 от 18.08.98.
Сдано в набор 20.05.10. Подписано в печать 11.08.10. Формат 60×84 1/8.
Бумага офсетная. Гарнитура SchoolBookC. Печать офсетная.
Усл. печ. л. 11,0. Уч.-изд. л. 14,0. Тираж 1000 экз. Заказ 267.

Оригинал-макет изготовлен в редакционно-издательском центре ГУАП.
190000, Санкт-Петербург, Б. Морская ул., 67.

Отпечатано с готовых диапозитивов в редакционно-издательском центре ГУАП.
190000, Санкт-Петербург, Б. Морская ул., 67.

ŋqdbyaY bE deqEbY3e d ɘɑɔqɔʎ ɟqdbYqyè

УДК 004.522

АВТОМАТИЧЕСКАЯ ОБРАБОТКА И СТАТИСТИЧЕСКИЙ АНАЛИЗ НОВОСТНОГО ТЕКСТОВОГО КОРПУСА ДЛЯ МОДЕЛИ ЯЗЫКА СИСТЕМЫ РАСПОЗНАВАНИЯ РУССКОЙ РЕЧИ

И. С. Кипяткова,
младший научный сотрудник
А. А. Карпов,
канд. техн. наук, старший научный сотрудник
Санкт-Петербургский институт информатики и автоматизации РАН

Описывается процесс автоматической обработки текстового корпуса, собранного из новостных лент ряда интернет-сайтов, для создания вероятностной n-граммной модели разговорного русского языка. Приводится статистический анализ данного корпуса, даются результаты по подсчету частоты появления различных n-грамм слов. Представлен обзор существующих типов статистических моделей языка.

Ключевые слова *— модель языка, текстовый корпус русского языка, автоматическая обработка текста.*

Введение

Для генерации грамматически правильных и осмысленных гипотез произнесенной фразы распознавателю речи необходима некоторая модель языка или грамматика, описывающая допустимые фразы. Процесс распознавания речи может быть представлен как поиск наиболее вероятной последовательности слов [1]:

$$W = \arg\max_{W} P(W \mid A) = \arg\max_{W} P(A \mid W)P(W),$$

где $P(A|W)$, $P(W)$ — вероятности появления гипотезы по оценке акустической и языковой модели соответственно.

Для многих языков (например, английского) разработаны методы создания моделей языка, которые позволяют повысить точность распознавания речи. Но эти методы не могут быть напрямую применены для русского языка из-за свободного порядка слов в предложениях и наличия большого количества словоформ для каждого слова.

Одной из наиболее эффективных моделей естественного языка является статистическая модель на основе n-грамм слов, цель которой состоит в оценке вероятности появления цепочки слов $W = (w_1, w_2, ..., w_m)$ в некотором тексте. n-граммы представляют собой последовательность из n элементов (например, слов), а n-граммная модель языка используется для предсказания элемента в последовательности, содержащей $n - 1$ предшественников. Эта модель основана на предположении, что вероятность какой-то определенной n-граммы, содержащейся в неизвестном тексте, можно оценить, зная, как часто она встречается в некотором обучающем тексте.

Вероятность $P(w_1, w_2, ..., w_m)$ можно представить в виде произведения условных вероятностей входящих в нее n-грамм [2]:

$$P(w_1, w_2, \ldots, w_m) = \prod_{i=1}^{m} P(w_i \mid w_1, w_2, \ldots, w_{i-1})$$

или аппроксимируя $P(W)$ при ограниченном контексте длиной $n - 1$:

$$P(w_1, w_2, \ldots, w_m) \cong \prod_{i=1}^{m} P(w_i \mid w_{i-n+1}, w_{i-n+2}, \ldots, w_{i-1}).$$

Вероятность появления n-граммы вычисляется на практике следующим образом:

$$P(w_i \mid w_{i-n+1}, \ldots, w_{i-1}) = \frac{C(w_{i-n+1}, \ldots, w_i)}{C(w_{i-n+1}, \ldots, w_{i-1})},$$

где C — количество появлений последовательности в обучающем корпусе.