УДК 004.522

И.С. Кипяткова, А.А. Карпов

РАЗРАБОТКА И ИССЛЕДОВАНИЕ СТАТИСТИЧЕСКОЙ МОДЕЛИ РУССКОГО ЯЗЫКА

Кипяткова И.С., Карпов А.А. **Разработка и исследование статистической модели русского языка**.

Аннотация. В статье описан процесс создания статистической модели русского языка для систем распознавания слитной речи. Дана характеристика собранного текстового корпуса, который сформирован из новостных лент ряда Интернет-сайтов электронных газет, проводится статистический анализ данного корпуса. На основе собранного текстового корпуса созданы униграммная, биграммная и триграммная модели русского языка. Для определения качества этих моделей использованы показатели энтропии и коэффициента неопределенности для этих моделей. Также в статье приведен обзор существующих подходов к созданию статистических моделей языка.

Ключевые слова: статистическая обработка текста, модель языка.

Kipyatkova I.S., Karpov A.A. **Development and Research of a Statistical Russian Language Model.**

Abstract. In the paper, the process of creation of a statistical Russian language model for continuous speech recognition systems is described. Characteristics of the collected corpus that consists of several news Internet sites of some on-line newspapers is given; a statistical analysis of this corpus is carried out. Unigram, bigram, and trigram Russian language models have been created on the base of the collected text corpus. For an estimation of quality of these models the entropy and perplexity parameters for these modelshave been computed. Also a survey of existing approaches for creation of statistical language models is given in the paper.

Keywords: statistical text processing, language model.

1. Введение. Для задачи распознавания речи с большим словарем необходима модель языка для генерации грамматически правильных и осмысленных гипотез произнесенной фразы. Одной из наиболее эффективных моделей естественного языка является статистическая модель на основе n-грамм, цель которой состоит в оценке вероятности появления цепочки слов $W = (w_1, w_2, \ldots, w_m)$ в некотором тексте. Фактически 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 $P(w_1, w_2, \ldots, w_m)$ можно представить в виде произведения условных вероятностей входящих в нее n-грамм [12]:

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

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 — число появлений последовательности в обучающем корпусе.

Наиболее простой моделью языка является нульграммная $(n=0)$ модель, которая предполагает, что каждое слово может следовать за любым другим словом. Тогда вероятность появления слова определяется как [10]:

$$P(w_i) = \frac{1}{|V|},$$

где $|V|$ — объем словаря.

Униграммная модель языка $(n=1)$ определяет вероятность появления i -го слова $P(w_i)$ в тексте. На практике обычно используются биграммная $(n=2)$, где определяется вероятность появления пар слов $P(w_i \mid w_{i-1})$, и триграммная $(n=3)$ модели языка, которая определяет вероятность появления троек слов $P(w_i \mid w_{i-2}, w_{i-1})$.

2. Разновидности статистических моделей языка. В данном разделе приведен обзор возможных вариантов построения моделей языка, основанных на статистическом анализе текста.

Модели, основанные на классах (Class-based models), используют функцию, которая отображает каждое слово w_i на класс $c_i : f : w_i \to f(w_i) = c_i$. В этом случае оценка условной вероятности может быть аппроксимирована по *n*-грамме класса [13]:

$$P(w_i \mid w_{i-n+1}, \ldots, w_{i-1}) = P(w_i \mid c_i) \cdot P(c_i \mid c_{i-1+1}, \ldots, c_{i-1}).$$

Функция отображения слова на класс может быть определена вручную с использованием некоторой морфологической информации (например, информации о части речи). Также существуют методы, которые помогают определить функцию отображения автоматически по текстовому корпусу.

Обычно n -граммные модели имеют ограниченный контекст $n = 2,3,4,5$, потому что с увеличением n очень быстро растет число параметров модели. Интервальные модели языка (*distance models*) помогают включить больший контекст, чем n -граммы, но коэффициент неопределенности модели остается того же порядка, как у n -грамм. Например, биграммная интервальная модель может быть задана следующим образом [13]:

$$P(w_i \mid w_{i-M+1},\ldots,w_{i-1}) = \sum_{m=1}^{M-1} \lambda_m P_m(w_i \mid w_{i-m}),$$

где M — предопределенное число моделей, $P_m(w_i \mid w_{i-m})$ — биграммная модель с пропуском $m-1$, λ_m — весовые параметры модели при условии $\sum_{m=1}^{M-1} \lambda_m = 1$.

Значение весовых коэффициентов λ_m определяется как зависимость от расстояния между словами w_i и w_{i-m} (с увеличением расстояния до слова весовой коэффициент уменьшается).

Триггерные модели (*trigger models*) — это другой тип моделей, которые моделируют взаимоотношение пар слов в более длинном контексте. В этом методе появление инициирующего слова в истории увеличивает вероятность появления другого слова, называемого целевым, с которым оно связано. Вероятность пар слов может быть определена следующим образом [13]:

$$P_{a\to b}(b \mid a \in h) = \frac{C(a \in h, b)}{C(a \in h)},$$

здесь a — это инициирующее слово; b — целевое слово, C — функция, определяющая подсчет события в текстовом корпусе; h — история некоторого ограниченного размера для слова b, т.е. те слова, которые предшествуют в тексте слову b.

Полная триггерная модель может быть определена следующим образом [13]:

$$P(w_i \mid w_M,\ldots,w_{i-1}) = \frac{1}{M}\sum_{m=1}^{M} \alpha(w_i, w_{i-m}), \quad \alpha(b,a) = \frac{P_{a\to b}(b \mid a \in h)}{\sum_w P_{a\to w}(w \mid a \in h)},$$

здесь M определяет длину цепочки слов в анализируемой истории h.

Упрощенной версией триггерных пар является кэш-модель (*cache model*). Она увеличивает вероятность появления слова в соответствии с тем, как часто данное слово употреблялось в истории, поскольку счи-

тается, что, употребив конкретное слово, диктор будет использовать это слово еще раз либо из-за того, что оно характерно для конкретной темы, либо потому что диктор имеет тенденцию использовать это слово в своем лексиконе. Кэш-модель можно рассматривать как простую n-граммную модель с вероятностями, вычисленными по предшествующей истории слов. Обычная униграммная кэш-модель может определяться как [13]:

$$P_C(w_i \mid h) = \frac{C(w_i, h)}{C(h)} = \frac{\sum_{j=i-D}^{i-1} I(w_i = w_j)}{\sum_{j=i-D}^{i-1} I(w_j \in V)},$$

где D — размер истории h, I — индикаторная функция, V — словарь модели языка.

Более развитые кэш-модели объединяются с убывающей функцией, которая моделирует следующее явление в языке: вероятность повторения слова уменьшается с увеличением промежутка от последнего появления слова в тексте:

$$P_{DC}(w_i \mid h) = \frac{\sum_{j=i-D}^{i-1} [I(w_i = w_j) \cdot d(i-j)]}{\sum_{j=i-D}^{i-1} d(i-j)},$$

где d — некоторая убывающая функция.

Другим типом модели языка является модель на основе набора тем (*topic mixture models*). Текстовый корпус вручную или автоматически делится на предопределенное число тем, и языковые модели создаются отдельно для каждой темы. Полная модель может определяться как [13]:

$$P_{TM}(w_i \mid h_i) = \sum_{j=1}^{M} \lambda_j \cdot P_j(w_i \mid h_i),$$

где M — число тем, P_j — модель темы j с весом модели λ_j.

Веса модели могут быть статическими или динамическими. Если используются динамические веса модели, то они устанавливаются для каждого слова w_i в зависимости от предшествующей истории. Такую модель можно называть адаптивной моделью.

Модели, основанные на частях слов, (*particle-based models*) используются для языков с богатой морфологией, например флективных

языков [13]. В этом случае слово w разделяется на некоторое число $L(w)$ частей (морфем) с помощью функции:

$$U : w \to U(w) = u^1, u^2, \ldots, u^{L(w)},\ u^i \in \Psi,$$

где Ψ — набор частей слова.

Разделение слов на морфемы можно производить двумя путями: при помощи 1) словарных и 2) алгоритмических методов [9].

Преимуществом алгоритмических методов является то, что они опираются лишь на анализ текста и не используют никаких дополнительных знаний, что позволяет анализировать текст на любом языке.

Преимущество словарных методов в том, что они позволяют получить правильное разбиение слов на морфемы, а не на псевдоморфемные единицы (как в алгоритмических методах), что может быть использовано далее на уровне пост-обработки гипотез распознавания фраз.

Хотя, по определению, n-граммные модели языка хранят только n слов, однако существуют модели, которые не ограничивают последовательности слов до определенного n, а вместо этого сохраняют различные последовательности разной длины. Такие модели называют n-граммами переменной длины (*varigrams*) [11]. По существу, они могут рассматриваться как n-граммные модели с большим n и такими принципами сокращения длины моделей, которые сохраняют только небольшой поднабор всех длинных последовательностей, встретившихся в обучающем тексте.

В [5] предлагается дальнодействующая триграммная модель, представляющая собой триграммную модель, в которой разрешены связи между словами, находящимися не только в пределах двух предыдущих слов, но и на большем расстоянии от предсказываемого слова. Лежащая в основе «грамматика» представляет собой множество пар слов, которые могут быть связаны вместе через несколько разделяющих слов.

В [6] предлагаются составные языковые модели и вводится понятие категорной языковой модели, в частности, категорных n-грамм. Каждому слову в словаре приписываются 15 атрибутов, определяющих грамматические свойства словоформы. Множество значений атрибутов определяет класс словоформы. Каждое слово в предложении рассматривается как его начальная форма и морфологический класс. В итоге грамматика разбивается на две составляющие: 1) изменяемую часть (основанную на морфологии) и 2) постоянную часть (основан-

ную на начальных формах слов), которая строится как n -граммная языковая модель.

В работе [2] для решения проблемы многозначности слов при автоматическом переводе с русского языка на латышский вместо биграмм используются синтаксические отношения и связи между парами элементов предложения. Из корпуса текстов латышского языка с помощью парсера выбираются синтаксически связанные пары слов. Определяется частота каждой уникальной пары, после чего вычисляется вероятность появления данной синтаксической пары.

3. Автоматическая статистическая обработка текстового корпуса для создания модели русского языка. Для обучения статистической модели языка необходим текстовый корпус большого объема. Существуют несколько текстовых корпусов русского языка, например, «Национальный корпус русского языка» (*www.ruscorpora.ru*) и «Корпус русского литературного языка» *(www.narusco.ru*), которые содержат в основном текстовый материал конца XX века. Эти корпуса включают в себя различные типы текстов: художественный, публицистический, научный, а также содержат в небольшом объеме и стенограммы устной речи. В работе [1] описан новостной корпус, собранный из примерно 2 тыс. источников средств массовой информации. Объем этого корпуса 7,3 млрд словоупотреблений.

Нами для создания модели языка собран и обработан новостной текстовый русскоязычный корпус, сформированный из новостных лент последних лет четырех Интернет-сайтов: *www.ng.ru* («Независимая газета»), *www.smi.ru* («СМИ.ru»), *www.lenta.ru* («LENTA.ru»), *www.gazeta.ru* («Газета.ru»). Он содержит тексты, отражающие срез современного состояния языка, в том числе и разговорного русского языка. Корпус может пополняться автоматически при обновлении сайтов в режиме *on line*, что позволяет оперативно добавлять новые появляющиеся в языке слова и переобучать модель языка с учетом новых текстовых данных. Естественный язык, будучи открытой системой, постоянно изменяется с изменением общественной жизни, развитием новых областей знаний, и *on line*-пополнение текстового корпуса позволяет учитывать изменения, происходящие в языке.

На базе собранного русскоязычного текстового корпуса создан частотный словарь объемом около 1 млн уникальных словоформ. Также для данного корпуса определена частота встречаемости различных би- и триграмм. Выполнена проверка соответствия текстового корпуса закону Ципфа (рис. 1). Закон Ципфа — эмпирическая закономерность распределения частоты слов естественного языка: ес-

ли все слова языка в достаточно большом осмысленном тексте упорядочить по убыванию частоты их использования, то частота слова в таком списке окажется приблизительно обратно пропорциональной его порядковому номеру [8]. Теоретически, для текстового корпуса график такой зависимости должен иметь вид, показанный на рис. 1 пунктирной линией. Полученный экспериментально график имеет незначительное отклонение от этой линии, но в целом можно заключить, что собранный корпус соответствует закону Ципфа.

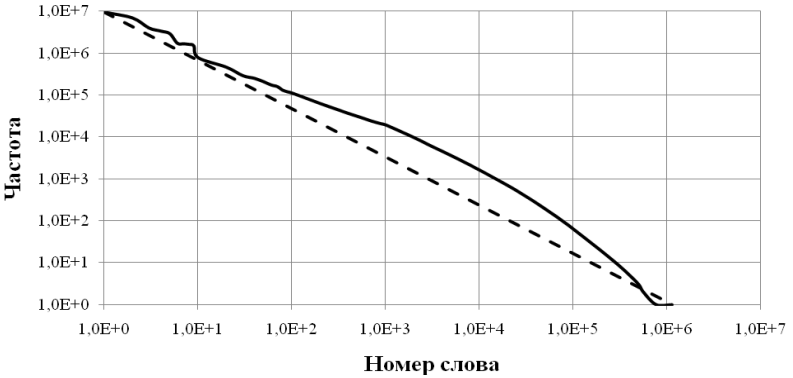

Рис. 1. Проверка соответствия текстового корпуса закону Ципфа.

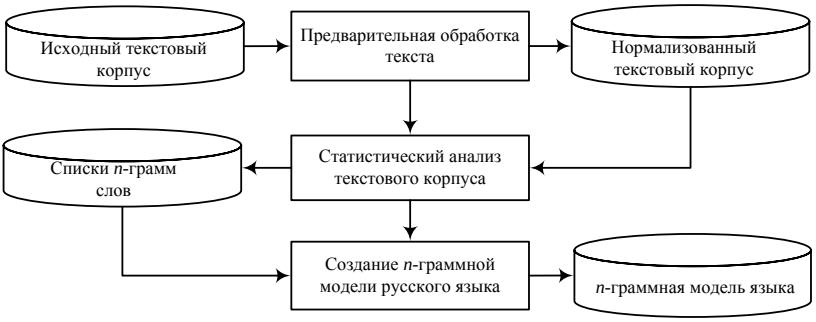

Рис.2. Диаграмма процесса создания модели языка.

Диаграмма процесса создания модели языка представлена на рис. 2. Автоматическая обработка собранного материала осуществля-

ется следующим образом [3]. Вначале происходит разбиение текстового массива на предложения, которые должны начинаться либо с заглавной буквы, либо с цифры. При этом учитывается, что в начале предложения могут стоять кавычки. Предложение заканчивается точкой, восклицательным или вопросительным знаком либо многоточием. Кроме того, при разделении текста на предложения учитывается, что внутри предложения могут стоять инициалы и фамилии. Формально это похоже на границу раздела двух предложений, поэтому если точка идет после одиночной заглавной буквы, то эта точка не будет считаться концом предложения. Предложения, содержащие прямую и косвенную речь, разделяются на отдельные предложения. Начало и конец предложения отмечаются знаками <s> и </s> соответственно. После разделения текстового материала на предложения выполняется его нормализация. Происходит удаление текста, написанного в любых скобках, удаление предложений, состоящих из пяти и меньшего количества слов (как правило это заголовки, составленные не по грамматическим правилам для полных предложений). Затем из текстов удаляются знаки препинания, расшифровываются общепринятые сокращения (например, обозначения размерности физических величин типа «см», «кг»). В словах, начинающихся с заглавной буквы, происходит замена заглавной буквы на строчную. Если все слово написано заглавными буквами, то замена не делается, так как это слово, вероятно, является аббревиатурой. На данный момент общий объем корпуса после его обработки — свыше 150 млн словоупотреблений (около 1 Гб данных).

Для автоматического распознавания речи необходимо иметь словарь фонематических транскрипций слов. Транскрипции для слов из собранного корпуса в основном создавались с помощью программного модуля, позволяющего создавать фонематические транскрипции слов автоматически [4]. Для генерации транскрипций модулю необходима база данных словоформ русского языка с отметкой ударения. В качестве таковой использовалась база данных словоформ русского языка, созданная путем объединения морфологических баз данных двух проектов, свободно доступных в Интернете:

1) Starling (*starling.rinet.ru*),

2) AOT (*www.aot.ru*).

Первая база данных содержит около 1,8 млн различных словоформ, но такой объем недостаточен для наших исследований. В этой базе данных для некоторых сложных слов проставлено также второстепенное ударение. Вторая база данных содержит свыше

2,2 млн словоформ. Однако в ней, в отличие от первой, отсутствует буква *ё* и информация о второстепенном ударении. Поэтому эти две базы данных были объединены.

Кроме того, в словарь добавлено более 40 тыс. словоформ из обрабатываемого тестового материала. Объем получившегося словаря превысил 2,38 млн различных словоформ. Вручную сделаны транскрипции для аббревиатур, а также для некоторых широко распространенных в наше время слов, заимствованных из английского языка.

На базе собранного русскоязычного текстового корпуса создан частотный словарь объемом около 1 млн уникальных словоформ. Статистическая модель языка создана с помощью программного модуля обработки и анализа текстов CMU (*Cambridge Statistical Language Modeling Toolkit*) [7]. Модель языка создавалась в несколько этапов. Вначале число биграмм составляло 22,85 млн, триграмм — 56,70 млн, число уникальных слов в текстах (словарь) — 937 тыс. Поскольку в обрабатываемом тексте присутствует достаточно большое число редких слов и слов с опечатками, при построении модели языка введен порог $K = 4$, т.е. n-граммы, у которых частота появления меньше 4, удалялись из модели языка. Затем для слов, которые использовались в этих моделях языка, были автоматически созданы транскрипции, а n-граммы со словами, для которых транскрипции не могли быть созданы автоматически (поскольку этих слов не было ни в словаре ударений, ни в списке аббревиатур, ни в списке иностранных слов), были удалены из модели. Однако из-за удаления некоторых n-грамм из модели языка появились слова, которые в модели не приводят к конечному результату (разрывают цепочку слов), поскольку встречаются в n-граммах не во всех позициях. Поэтому модель языка была также сокращена путем удаления n-грамм, содержащих такие слова. В результате в конечной биграммной модели число уникальных словоформ составило 182 тыс., число биграмм — 3,17 млн, в триграммной модели число уникальных словоформ — 126 тыс., триграмм — 3,54 млн. В таблице представлено число n-грамм, которое получалось на этих этапах.

Объем словаря и число *n*-грамм на этапах создания модели языка

Этап создания модели	Биграммная модель		Триграммная модель	
	Объем словаря, тыс. слов	Число n-грамм, млн	Объем словаря, тыс. слов	Число n-грамм, млн

До фильтрации n -грамм	937	22,85	937	56,70
После удаления n -грамм с порогом $K = 4$	273	3,35	201	3,90
После удаления n -грамм со словами без транскрипций	234	3,23	176	3,64
Модель после фильтрации n-грамм со словами, встречающихся не во всех позициях в n -граммах	182	3,17	126	3,54

4. Вычисление энтропии и коэффициента неопределенности созданной модели языка. Для тестирования созданной модели языка собран корпус меньшего объема, содержащий текстовый материал новостного сайта *www.fontanka.ru* («Фонтанка.ru»). На этом тестовом корпусе вычислены энтропия и коэффициент неопределенности (*perplexity*) статистической модели языка. По определению, информационная энтропия — мера хаотичности информации, неопределенность появления какого-либо символа первичного алфавита. При отсутствии информационных потерь она численно равна количеству информации, приходящейся на один символ передаваемого сообщения. Поскольку тексты на естественном языке могут рассматриваться в качестве информационного источника, энтропия вычисляется по следующей формуле [11]:

$$H = -\lim_{m\to\infty}\frac{1}{m}\sum_{w_1,w_2,\ldots,w_m}\left(P(w_1,w_2,\ldots,w_m)\log_2 P(w_1,w_2,\ldots,w_m)\right).$$

Это суммирование делается по всем возможным последовательностям слов. Но поскольку язык является эргодичным источником информации [11], выражение для вычисления энтропии имеет следующий вид:

$$\hat{H} = -\frac{1}{m}\log_2 P(w_1,w_2,\ldots,w_m).$$

Коэффициент неопределенности является параметром, по которому оценивается число n -граммных моделей языка, и вычисляется следующим образом [11]:

$$PP = 2^{\hat{H}} = \hat{P}(w_1,w_2,\ldots,w_m)^{-\frac{1}{m}},$$

где $\hat{P}(w_1,w_2,\ldots,w_m)$ — вероятность последовательности слов $w_1,w_2,\ldots,w_m$.

Коэффициент неопределенности показывает, сколько в среднем различных наиболее вероятных слов может следовать за данным словом. Для униграммной модели коэффициент неопределенности равен 5909,08, энтропия — 12,53 бит/слово, для биграммной модели коэффициент неопределенности равен 1579,29, энтропия — 10,63 бит/слово, для триграммной модели коэффициент неопределенности равен 1031,77, энтропия — 10,01 бит/слово. При этом относительное число внесловарных слов при использовании униграммной и биграммной модели составило 7,97 %, при использовании триграммной модели — 10,88 %. Полученные значения достаточно велики. Например, для английского языка при объеме словаря в 200 тыс. слов, коэффициент неопределенности для биграмм равен 232 [14], при этом энтропия будет приблизительно равна 7,9 бит/слово, а относительное число внесловарных слов составляет 0,31 % для тестового корпуса объемом 1,12 млн. слов.

5. Заключение. Текстовый материал для статистической обработки взят из Интернет-сайтов четырех электронных газет. Таким образом, корпус, предназначенный для создания модели языка, основывается на текстах с большим числом стенограмм выступлений и прямой речи, отражающих особенности современного языка, а не на литературных текстах, которые крайне далеки от разговорной речи. Разработанная методика сбора текстового материала позволяет при обновлении Интернет-сайтов оперативно дополнять текстовый корпус и затем переобучать модель языка в режиме *on line*, учитывая тем самым изменения, происходящие и в самом языке, и в контексте текущих событий. Однако использование Интернет-материалов имеет ряд недостатков, главным из которых является наличие в текстах опечаток. Кроме того, в таких текстах присутствует много имен собственных, большинство из которых в разговорной речи встречается редко. Из-за этого возрастает объем созданных в результате обработки текста n-грамм.

На основе новостного текстового корпуса созданы уни-, би- и триграммная модели русского языка. Однако проведенный анализ показывает, что стандартные методы создания моделей языка не подходят для русского языка, поскольку в нем очень велико соотношение уникальных слов к размеру текстового корпуса. Для решения данной проблемы целесообразно создавать модель языка, основываясь на начальных формах слов или используя основы слов. Это позволит сократить объем словаря распознавателя и списков n-грамм.

Литература

1. *Баглей С.Г., Антонов А.В., Мешков В.С., Суханов А.В.* Статистические распределения слов в русскоязычной текстовой коллекции // Материалы междунар. конф. «Диалог 2009». Москва. 2009. С. 13–18.
2. *Горностай Т., Васильев А., Скадиньш Р., Скадиня И.* Опыт латышско↔русского машинного перевода // Материалы междунар. конф. «Диалог 2007». Москва. 2007. С. 137–146.
3. *Кипяткова И.С., Карпов А.А.* Автоматическая обработка и статистический анализ новостного текстового корпуса для модели языка системы распознавания русской речи // Информационно-управляющие системы. 2010. № 4(47). С. 2–8.
4. *Кипяткова И.С., Карпов А.А.* Модуль фонематического транскрибирования для системы распознавания разговорной русской речи // Искусственный интеллект. 2008. № 4. С. 747–757.
5. *Протасов С.В.* Вывод и оценка параметров дальнодействующей триграммной модели языка // Материалы междунар. конф. «Диалог 2008». Москва. 2008. С. 443–449.
6. *Холоденко А.Б.* О построении статистических языковых моделей для систем распознавания русской речи // Интеллектуальные системы. 2002. Т. 6, вып. 1–4. С. 381–394.
7. *Clarkson P., Rosenfeld R.* Statistical language modeling using the CMU-Cambridge toolkit // Proc. of EUROSPEECH. Rhodes. Greece. 1997. P. 2707–2710.
8. *Gelbukh A., Sidorov G.* Zipf and Heaps Laws' Coefficients Depend on Language // Proc. CICLing-2001, Conf. on Intelligent Text Processing and Computational Linguistics. Mexico City. Lecture Notes in Computer Science № 2004. 2001. Springer-Verlag. P. 332–335.
9. *Kurimo M., Hirsimäki T., Turunen V.T., Virpioja S. et al.* Unsupervised decomposition of words for speech recognition and retrieval // Proc. of 13th Intern. Conf. SPECOM'2009. St. Petersburg. 2009. P. 23–28.
10. *Merkel A., Klakow D.* Improved Methods for Language Model Based Question Classification // Proc. of 8th Interspeech Conf. Antwerpen. 2007. P. 322–325.
11. *Moore G.L.* Adaptive Statistical Class-based Language Modelling. PhD thesis. Cambridge University. 2001. 193 p.
12. *Rabiner L., Juang B.-H.* Fundamentals of Speech Recognition. Prentice Hall, 1995. 507 p.
13. *Vaičiūnas A.* Statistical Language Models of Lithuanian and Their Application to Very Large Vocabulary Speech Recognition. Summary of Doctoral Dissertation. Kaunas: Vytautas Magnus University, 2006. 35 p.
14. *Whittaker E.W.D.* Statistical Language Modelling for Automatic Speech Recognition of Russian and English. PhD thesis. Cambridge University. 2000. 140 p.

Кипяткова Ирина Сергеевна — м. н. с. лаборатории речевых и многомодальных интерфейсов Учреждения Российской академии наук Санкт-Петербургского института информатики и автоматизации РАН (СПИИРАН). Область научных интересов: автоматическое распознавание речи, статистические модели языка. Число научных публикаций — 15. kipyatkova@iias.spb.su; СПИИРАН, 14-я линия В.О., д. 39, Санкт-Петербург, 199178, РФ; р.т. +7(812)328-7081, факс +7(812)328-7081. Научный руководитель — канд. техн. наук А.А. Карпов.

Kipyatkova Irina Sergeevna — junior researcher, Laboratory of Speech and Multimodal Interfaces St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences (SPIIRAS). Research interests: automatic speech recognition statistical language models. The number of publications — 15. kipyatkova@iias.spb.su; SPIIRAS, 39, 14th Line

V.O., St. Petersburg, 199178, Russia; office phone +7(812)328-7081, fax +7(812)328-7081. Scientific adviser — PhD A.A. Karpov.

Карпов Алексей Анатольевич — канд. техн. наук**,** с. н. с. лаборатории речевых и многомодальных интерфейсов Учреждения Российской академии наук Санкт-Петербургского института информатики и автоматизации РАН (СПИИРАН). Область научных интересов: автоматическое распознавание речи, многомодальные интерфейсы, аудиовизуальное распознавание речи. Число научных публикаций — 100. karpov@iias.spb.su; СПИИРАН, 14-я линия В.О., д. 39, Санкт-Петербург, 199178, РФ; р.т. +7(812)328-7081, факс +7(812)328-7081.

Karpov Alexey Anatolyevich — PhD, senior researcher, Laboratory of Speech and Multimodal Interfaces St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences (SPIIRAS). Research interests: automatic speech recognition, multimodal interfaces, audio-visual speech recognition. The number of publications — 100. karpov@iias.spb.su; SPIIRAS, 39, 14th Line V.O., St. Petersburg, 199178, Russia; office phone +7(812)328-7081, fax +7(812)328-7081.

Поддержка исследований. Данное исследование поддержано Министерством образования и науки РФ в рамках ФЦП «Научные и научно-педагогические кадры инновационной России» (госконтракты № 14.740.11.0357, П2579, П2360, П876) и международным фондом «Научный Потенциал» (договор № 201).

Рекомендовано лабораторией речевых и многомодальных интерфейсов, заведующий лабораторией д-р техн. наук, доц. А.Л. Ронжин.
Статья поступила в редакцию 15.11.2010.

РЕФЕРАТ

Кипяткова И.С., Карпов А.А. **Разработка и исследование статистической модели русского языка.**

Модель языка необходима для систем распознавания речи с большим словарем. Одной из наиболее эффективных моделей естественного языка является статистическая модель на основе n-грамм. В статье описаны некоторые разновидности статистических моделей языка: интервальные, триггерные, модели, основанные на классах, кэш-модели, модели на основе набора тем, модели, основанные на частях слов, n-граммы переменной длины.

В статье представлен процесс создания n-граммной модели русского языка. Для создания модели языка собран и обработан новостной текстовый корпус, который был сформирован из Интернет-сайтов ряда электронных газет. Таким образом, корпус основывается на текстах с большим числом стенограмм выступлений и прямой речи, отражающих особенности современного языка. Собранный корпус был автоматически обработан. Текст был разделен на предложения, удалены предложения, состоящие из пяти слов и меньше, также был удален текст, написанный в скобках. Произведена расшифровка общепринятых сокращений, удалены знаки препинания. Объем корпуса после его обработки составил более 150 млн словоупотреблений, при этом число уникальных словоформ — около 1 млн. Определено, что собранный корпус соответствует закону Ципфа. Корпус был статистически обработан, созданы униграммная, биграммная и триграммная модели языка. Поскольку в обрабатываемом тексте присутствует достаточно большое число редких слов и слов с опечатками, при построении модели языка был введен порог, и n-граммы с частотой появления меньше 4 были удалены из модели языка. Затем для слов, которые использовались в этих моделях языка, были автоматически созданы транскрипции; n-граммы со словами, для которых транскрипции не могли быть созданы автоматически, удалялись из модели языка. Затем из модели языка были удалены n-граммы со словами, которые встречаются в n-граммах не во всех позициях и таким образом в модели не приводят к конечному результату. В итоге в конечной биграммной модели число уникальных словоформ составляет 182 тыс., количество биграмм — 3,17 млн, в триграммной модели число уникальных словоформ — 126 тыс., триграмм — 3,54 млн.

Для тестирования созданной модели языка собран текстовый корпус, содержащий материал сайта электронной газеты «Фонтанка.ru». Для этого корпуса определены энтропия и коэффициент неопределенности созданной модели языка. Для униграммной модели коэффициент неопределенности равен 5909,08, энтропия — 12,53 бит/слово, для биграммной коэффициент неопределенности равен 1579,29, энтропия — 10,63 бит/слово, для триграммной коэффициент неопределенности равен 1031,77, энтропия — 10,01 бит/слово. При этом относительное число новых слов при использовании уни- и биграммной модели равно 7,97 %, при использовании триграммной модели — 10,88 %.

SUMMARY

Kipyatkova I.S., Karpov A.A. **Development and Research of a Statistical Russian Language Model.**

A language model is necessary for any large vocabulary speech recognition system. One of the most effective natural language models is a statistical model based on n-grams. In the paper, some kinds of statistical language models are described: distance models, trigger models, class-based models, cache models, topic mixture models, particle-based models, varigrams.

In the paper, a process of creation of n-gram Russian language model is described. A news text corpus that consists of Internet sites of some on-line papers has been collected and processed for creation of the language model. Thus, the corpus is based on some texts with lots of shorthand reports and direct speech describing properties of contemporary spoken language. The collected corpus has been processed automatically. The text was divided into sentences; sentences consisting of less than six words were deleted, a text written in any brackets was also deleted. Abbreviations expansion was carried out, any punctuation marks were deleted. After this processing, the corpus has above 150 M word usages; a quantity of unique word-forms is about 1 million. It was determined that the collected corpus agrees well with the Zipf's law. The corpus has been statistically processed as well; unigram, bigram, and trigram language models were created. As far as in the text processing text there are lots of rare words and words with mistakes, a threshold was introduced while model creating, and n-grams, the frequency of appearance of which was less than 4, were deleted from the language model. Then transcriptions were automatically created for all the words in these language models; n-grams with words, transcriptions for which could not be created automatically, were removed from the language model. Then n-grams with words that appear not in all positions in n-grams and thus do not lead to the result, were deleted. As a result in the final bigram model, number of unique word-forms is 182 K, number of bigrams is 3.17 M, in the trigram model number of unique word-forms is 126 K, and number of trigrams is 3.54 M.

A corpus containing text material of on-line newspapers (“Fontanka.ru”) has been collected to test the language model. Entropy and perplexity of the created language model have been calculated for this corpus. For the unigram language the model perplexity is 5909.08, entropy equals 12.53 bit/word, for the bigram model the perplexity is 1579.29, entropy equals 10.63 bit/word, for the trigram model perplexity is 1031.77, entropy equals 10.01 bit/word. At the same time, a relative number of new (out-of-vocabulary) words is 7.97 % for the unigram and bigram model, and 10.88 % — for the trigram model.