OSTIS-2011
(Open Semantic Technologies for Intelligent Systems)

УДК 004.522

ОПЫТ РАЗРАБОТКИ МОДЕЛИ РАСПОЗНАВАНИЯ РУССКОЙ РЕЧИ СО СВЕРХБОЛЬШИМ СЛОВАРЕМ

И.С. Кипяткова (*kipyatkova@iias.spb.su*)
А.А. Карпов (*karpov@iias.spb.su*)
Учреждение Российской академии наук Санкт-Петербургский институт информатики и автоматизации РАН, г. Санкт-Петербург, Россия

В статье описывается процесс создания статистических моделей русского языка для систем распознавания слитной речи. Модели языка были созданы по текстовому корпусу, сформированному из новостных лент ряда интернет-сайтов электронных газет, была проведена автоматическая статистическая обработка текстового корпуса. Также в статье представлены результаты экспериментов по распознаванию слитной речи со сверхбольшим словарем с применением *n*-граммных моделей языка.

Ключевые слова: модель языка, распознавание слитной русской речи, сверхбольшой словарь, статистическая обработка текста.

Введение

Одной из основных нерешенных проблем в области речевых исследований является автоматическое стенографирование или распознавание слитной разговорной речи. Согласно принятой сейчас в мире классификации, малым словарем распознавания считается словарь в единицы и десятки слов [Benesty et al., 2008]. Задач и приложений, где используется малый словарь распознавания, достаточно много: распознавание последовательностей цифр, номеров телефонов; системы речевого командного управления и т.д. Средний распознаваемый словарь содержит сотни слов. Такого словаря достаточно для большинства диалоговых или запросно-ответных систем. Большой словарь содержит тысячи и десятки тысяч слов, такие системы распознавания могут использоваться в автоматизированных справочных системах или системах диктовки текста в ограниченной предметной области (для аналитических языков). Словарь размером в сотни тысяч и миллионы слов считается сверхбольшим, он позволяет реализовывать системы стенографирования текста (включая синтетические языки) [Hori et al., 2006].

Мировых исследований, посвященных разработке систем распознавания речи со сверхбольшим словарем, относительно немного. Это связано с тем, что для многих языков такой словарь был бы избыточным. Так в работе [Whittaker, 2000] показано, что при размере словаря в 65 тыс. слов английского языка, количество внесловарных слов (out-of-vocabulary words) составляет 1,1 %. Для флективных же языков, к числу которых относится и русский, из-за наличия большого числа словоформ для каждой парадигмы слова объем словаря распознавания и количество существующих внесловарных слов возрастают на порядок по сравнению с аналитическими языками.

Для задачи распознавания речи с большим и сверхбольшим словарем необходима модель языка для генерации грамматически правильных и семантически связанных гипотез произнесенной фразы. Одной из наиболее эффективных моделей естественного языка является статистическая модель на основе *n*-грамм. В статье представлены результаты распознавания слитной русской речи со сверхбольшим словарем с применением *n*-граммных моделей языка при *n*, равном от 0 до 3.

1. Создание статистической модели русского языка

Для создания модели языка нами был 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. Пополнение этого корпуса может осуществляться автоматически при обновлении сайтов в режиме он-лайн, что позволяет оперативно добавлять новые появляющиеся в языке слова и переобучать модель языка с учетом новых текстовых данных.

Диаграмма процесса создания модели языка представлена на рисунке 1. Автоматическая обработка собранного материала осуществляется следующим образом [Кипяткова и др., 2010]. Вначале происходит разбиение текстового массива на предложения, при этом предложения, содержащие прямую и косвенную речь, разделяются на отдельные предложения. Начало и конец предложения отмечаются знаками <s> и </s> соответственно. Удаляются повторяющиеся предложения. Затем 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, «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 составляет свыше 110 млн словоупотреблений (около 750 Мб данных).

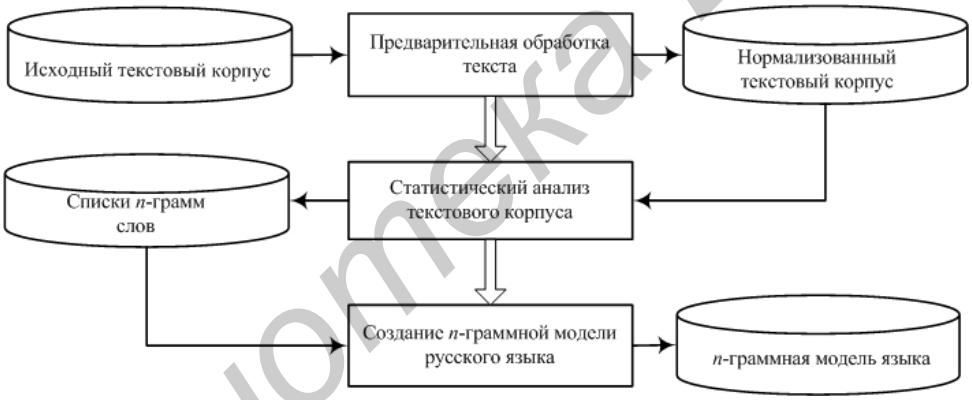

Рисунок 1 — Диаграмма процесса создания модели языка

На базе собранного русскоязычного текстового корпуса был создан частотный словарь, размер которого составляет около одного миллиона уникальных словоформ. Статистическая модель языка была создана с помощью программного модуля обработки и анализа текстов CMU (Cambridge Statistical Language Modeling Toolkit) [Clarkson et al., 1997]. Модель языка создавалась в несколько этапов. Вначале количество биграмм составляло 22,7 млн, триграмм — 56,4 млн, количество 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 был введен порог K, то есть n-граммы, у которых частота появления меньше K, удалялись из модели языка. Для биграммной модели языка был утсановлен порог K=2. При создании триграммной модели языка был выбран порог K=3, поскольку при меньшем пороге из-за большого числа триграмм вероятность некоторых из них была настолько мала, что возникали ошибки при округлении, в результате чего сумма вероятностей оказывалась больше 1. Затем для слов, которые использовались в этих моделях языка, были автоматически созданы транскрипции [Кипяткова и др., 2008]. n-граммы со словами, для которых транскрипции не могли быть созданы автоматически, были удалены из модели языка. 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 количество уникальных словоформ составило 208 тыс., количество биграмм — 6,01 млн, в триграммной модели количество уникальных словоформ — 76 тыс., триграмм — 3,43 млн.

2. Результаты распознавания речи с использованием *n*-граммных моделей языка

Для распознавания слитной русской речи использовался разработанный в СПИИРАН декодер SIRIUS [Ronzhin et al., 2007]. В качестве фонетических единиц при распознавании речи применялись контекстно-зависимые фоны (трифоны). Запись обучающего и тестового речевого корпуса для системы производилась с частотой дискретизации 44 кГц, 16 бит на отсчет, моно, отношение сигнал/шум больше 35 дБ. Для обучения системы распознавания речи были использованы 300 вручную размеченных фраз из речевого корпуса [Jokisch et al., 2009]. Система была обучена на записях одного диктора и является, таким образом, дикторозависимой. Для тестирования системы были записаны 100 слитно произнесенных фраз, состоящих из 1068 слов (7191 символов); фразы взяты из материалов интернет-газеты «Фонтанка.ru» (www.fontanka.ru). В текстовом корпусе, используемом для тестирования, количество биграмм, присутствующих в модели языка, составило 83,58 %, триграмм — 35,83 %, при этом для словаря объемом 208 тыс. слов относительное количество внесловарных слов было равно 0,75 %, для словаря объемом 76 тыс. слов — 4,87 %. Для этого тестового корпуса вычислены величины энтропии и коэффициента неопределенности (perplexity) статистической модели языка [Moore, 2001]. Для униграммной модели величина коэффициента неопределенности составила 5493,11, энтропии — 12,42 бит/слово, для биграммной модели коэффициент неопределенности равен 776,67, энтропия — 9,60 бит/слово, для триграммной модели коэффициент неопределенности равен 452,14, энтропия — 8,82 бит/слово. Полученные значения являются достаточно большими. Например, для английского языка при размере словаря в 200 тыс. слов, коэффициент неопределенности для биграммной модели равен 232 [Whittaker, 2000], при этом энтропия будет приблизительно равна 7,9 бит/слово.

Результаты распознавания слов и символов (под символом понимаются буква и знак пробела) с применением различных моделей языка представлены в таблицах 1 и 2. Методика оценивания результатов распознавания описана в [Young et al., 2009]. Для нульграммной, униграммной и биграммной модели использовался одинаковый словарь объемом в 208 тыс. слов, для триграммной модели объем словаря был 75 тыс. слов. При использовании нульграммной модели (то есть при распознавании без модели языка) точность распознавания слов оказалась отрицательной и равной -20,97 %, это связано с тем, что количество вставленных слов было больше чем количество правильно распознанных. Использование униграммной модели позволило повысить точность распознавания слов до 30,06 %. Наилучшие результаты были достигнуты при применении биграммной модели, где точность распознавания слов составила 36,89 %. При применении триграммной модели точность распознавания упала до 24,72 %. Снижение точности при распознавании с триграммной моделью языка связано с малым размером словаря, используемым в данной модели, в результате чего в тестовых фразах количество триграмм, присутствующих в модели языка, было также мало (35,83 %). Однако точность распознавания с использованием триграммной модели была значительно выше, чем при распознавании речи без применения модели языка, но с большим словарем.

Таблица 1 — Результаты распознавания слов с использованием различных моделей языка

Тип модели	Точность распознавания, %	Количество правильно распознанных слов	Количество удаленных слов	Количество замененных слов	Количество вставленных слов
0-граммная	-20,97	192	23	853	416
1-граммная	30,06	328	244	496	7
2-граммная	36,89	559	43	466	165
3-граммная	24,72	270	373	425	6

Таблица 2 — Результаты распознавания символов с использованием различных моделей языка

Тип модели	Точность распознавания, %	Количество правильно распознанных символов	Количество удаленных символов	Количество замененных символов	Количество вставленных символов
0-граммная	54,47	4745	986	1460	828
1-граммная	67,47	5044	1262	885	192
2-граммная	73,70	5660	722	809	360
3-граммная	60,87	4571	1615	1005	194

Заключение

Статистические модели языка были созданы по текстовому корпусу, сформированному из новостных интернет-сайтов четырех электронных газет. Таким образом, этот корпус содержит тексты с большим количеством стенограмм выступлений и прямой речи, отражающих особенности современного языка, а не на литературных текстах, которые крайне далеки от разговорной речи. Были проведены эксперименты по дикторозависимому распознаванию слитно произнесенных фраз с применением нульграммной, униграммной, биграммной и триграммной моделей языка. Наивысшая точность распознавания (36,89 %) была достигнута при применении биграммной модели. К сожалению, проведенные эксперименты не позволяют судить об эффективности применения триграммной модели, поскольку размер словаря для нее был существенно меньше. В дальнейшей работе мы намереваемся модифицировать базовую *n*-граммную модель с учетом особенностей русского языка. Мы планируем исследовать модели языка, строящиеся на основе начальных форм слов и основах слов, а также модели языка, использующие грамматические морфологические показатели слов, полученные в результате семантико-сентаксического разбора предложений.

Данное исследование поддержано фондом РФФИ (проекты № 08-08-00128, 09-07-91220-СТ), Министерством образования и науки РФ в рамках ФЦП «Научные и научно-педагогические кадры инновационной России» (госконтракты № П2579, П2360) и Комитетом по науке и высшей школе Правительства Санкт-Петербурга.

Библиографический список

[Кипяткова и др., 2008] Кипяткова И.С., Карпов А.А. Модуль фонематического транскрибирования для системы распознавания разговорной русской речи // Искусственный интеллект. Донецк. Украина. № 4. 2008. С. 747-757.

[Кипяткова и др., 2010] Кипяткова И.С., Карпов А.А. Автоматическая обработка и статистический анализ новостного текстового корпуса для модели языка системы распознавания русской речи // Информационно-управляющие системы. СПб.: СПбГУАП, № 4(47). 2010. С. 2-8.

[Benesty et al., 2008] Benesty J., Sondhi M., Huang Y. (eds.) Springer Handbook of Speech Processing. Springer, 2008, 1176 p.

[Clarkson et al., 1997] Clarkson P., Rosenfeld R. Statistical language modeling using the CMU-Cambridge toolkit // Proc. of EUROSPEECH. Rhodes. Greece. 1997. pp. 2707–2710.

[Hori et al., 2006] Hori, T., Nakamura A. An extremely-large-vocabulary approach to named entity extraction from speech. Proceedings of ICASSP'2006, Toulouse, France, 2006.

[Jokisch et al., 2009] Jokisch O., Wagner A., Sabo R., Jaeckel R., Cylwik N., Rusko M., Ronzhin A., Hoffmann R. Multilingual Speech Data Collection for the Assessment of Pronunciation and Prosody in a Language Learning System //Proc. of 13-th International Conference SPECOM'2009. St. Petersburg. 2009. pp. 515–520.

[Moore, 2001] Moore G.L. Adaptive Statistical Class-based Language Modelling. PhD thesis. Cambridge University. 2001. 193 p.

[Ronzhin et al., 2007] Ronzhin A.L., Karpov A.A. Russian Voice Interface // Pattern Recognition and Image Analysis. (Advances in Mathematical Theory and Applications). 2007. Т. 17. № 2. С. 321-336.

[Whittaker, 2000] Whittaker E.W.D. Statistical Language Modelling for Automatic Speech Recognition of Russian and English. PhD thesis. Cambridge University. 2000, 140 p.

[Young et al., 2009] Young S. et al. The HTK Book (for HTK Version 3.4). Cambridge, UK, 2009.