АНАЛИЗ ВРЕМЕННЫХ РЯДОВ И СИМВОЛЬНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ
(Вычислительные системы)
1991 год Выпуск 141

УДК 621.391

ФОРМИРОВАНИЕ УНИВЕРСАЛЬНОГО ОБУЧАЮЩЕГО ТЕКСТА ДЛЯ СИСТЕМ РАСПОЗНАВАНИЯ РЕЧИ

В.М.Величко, Н.В.Саломатина, Н.В.Чукалина, Л.С.Юдина

В в е д е н и е

В статье рассматриваются вопросы формирования универсального обучающего текста для использования в системах распознавания речи, а также для некоторых других приложений. Статья является продолжением работы [1].

Для оценки матриц переходов в алгоритмах распознавания речи с использованием модели скрытых марковских процессов желательно располагать обучающей выборкой, содержащей *все* двухфонемные сочетания (биграммы), встречающиеся в *русской слитной* речи, для работы впоследствии с любыми словарями [2]. Эта задача решена для английской речи [3]. В работе [3] используются выборки длительностью 20 мин и 15 сек, дающие достаточно хорошие оценки требуемых матриц переходов для марковской модели.

Первоначальной целью работы являлось составление и оптимизация (с использованием ЭВМ) текста, где каждая возможная в языке биграмма встретилась бы, по крайней мере, пять раз в разных контекстах. В процессе выполнения работы пришлось внести существенные коррективы в первоначальный план, о чем будет сказано ниже. Цель настоящей работы скромнее - добиться хотя бы

одноразовой встречаемости каждой биграммы при минимальном (по возможности) объеме обучающего текста.

Ниже обсуждаются некоторые методы и средства формирования заданного текста .

1. Требования к универсальному тексту

Основные требования к тексту формулировались априорно разработчиками систем распознавания речи и заключались в следующем:

- текст состоит из отдельных фраз;
- внутри фразы слитная (без пауз) осмысленная речь;
- простота синтаксиса и легкость произнесения слов;
- во фразе в среднем пять слов;
- минимальная частота встречаемости каждой биграммы - 5 раз;
- небольшой объем текста (желательно, не более 20 минут непрерывного звучания).

Требования обуславливались следующими соображениями. Пофразовое произнесение обеспечивает необходимую степень редукции фонем в слитной речи, что позволяет набирать статистику в естественной речи, избегая искусственного "полного стиля". Осмысленность фраз (при разумном понимании осмысленности) делает произнесение более естественным. Требование отсутствия пауз обусловлено необходимостью получения сочетаний фонем на стыках слов, где встречаются наиболее редкие биграммы, составляющие значительную по количеству часть всех возможных биграмм. Простота и легкость произнесения фраз - естественное эргономическое требование к работоспособной системе. Средняя длина фразы (5 слов) обеспечивает запас дыхания на слитное произнесение. Требование встречаемости каждой биграммы обеспечивает полноту обучающей выборки (генеральную совокупность),а 5-кратный повтор - минимальную представительность генеральной сово -

купности. Объем текста задавался из двух соображений. Во-первых, по аналогии с английским вариантом [3]. Во-вторых, исходя из предполагаемого количества разрешенных биграмм русского языка, которое оценивалось примерно в 1600 единиц [1,4]. Таким образом, при 5-кратном повторении каждой биграммы для реальных сочетаний будет порядка 8000 фонем, а с учетом неравномерной их повторяемости - примерно в полтора раза больше,т.е. порядка 12000, что соответствует 20 минутам речи [1].

2. Формирование текста

Специалистам известно, что частота встречаемости фонем и сочетаний фонем в речи далеко не одинакова. Интервал между полярными ее значениями достигает четырех порядков. Так, например, в проведенных нами ранее исследованиях на 1,5 млн. фонем биграмма [нъ] встретилась в текстах около 23 тыс. раз, а биграмма [р'к] - 3 раза, биграмма [л'х] - 2 раза [3].

В стремлении преодолеть закон столь неравномерного распределения языковых единиц, сбалансировать частоту встречаемости биграмм в целях сокращения объема текста заключается основная трудность работы.

В нарушение поставленного условия простоты и легкости произнесения текста оказалось необходимым привлечение значительной части иноязычных по происхождению слов,поскольку именно они содержат редкие биграммы.

В произвольном русском транскрибированном тексте объемом в 1,5 млн.знаков *внутри слова* реально встречается около 1600 биграмм из 2500 возможных [4]. Достоверная статистика биграмм, встречающихся на границах слов,авторам не известна. Априори мы исходим из предположения о том, что возможны сочетания всех фонем со всеми, кроме сочетаний, запрещенных языковыми законами, - таковых около 1000. Точнее, из 3135 возможных биграмм не противоречат языковым законам и, следовательно, могут быть реализованы 2013.

Наш инвентарь фонем включает 56 единиц:

- 6 гласных (а́,о́,у́,э́,ы́,и́);

- 36 согласных (бб',вв',гг',дд',ж,зз',кк',лл',мм',нн',пп',рр',сс',тт',фф',хх',ц,ч,ш,$\bar{ш}$',j);

- 7 аллофонов гласных (ъ, ʌ,и^{е},ь,и,ы,у).

Кроме того, как уже говорилось, слова во фразах произно - сятся без пауз, и на межсловных границах в числе других процессов происходит ассимиляция по звонкости между согласными. В связи с этим нами предусмотрено озвончение переднеязычных [ц, ч,$\bar{ш}$'] и заднеязычного [х]. В результате инвентарь фонем увеличился еще на 4 аллофона согласных, а именно:

- ц → $\widehat{дз}$ (z) [ʌт'э́zгъвʌр'и́л] "Отец говорил";

- ч → $\widehat{д'ж'}$ (g) [кл'у́gзʌбы́л] "Ключ забыл";

- $\bar{ш}$' → $\bar{ж}$' ($\bar{ж}$') [бо́р$\bar{ж}$'зб'и^{е}жа́л] "Борщ сбежал";

- х → γ(Y) [ду́Yза́м'ьр] "Дух замер".

Особое внимание уделено односложным словам с безударными [о,е,ə], например: "сто один" - [стøʌд'и́н], "тем самым" - [т'ем са́мым], "а также" - [ʌта́гжъ]. Таким образом, в состав фонем вошли безударные [о] и [е], обозначенные как [ø] и [е]; безударный [а] совпадает в обозначении с [ʌ]. Включен в состав фонем также символ [#], обозначающий начало и конец фразы. В сочетании с начальными и конечными фонемами он образует самостоятельные биграммы. Таким образом, в нашу задачу входило: реализовать во фразах свыше 2 тысяч биграмм из 3135 возможных.

Транскрипция текста производилась по алгоритму, описанному в [5], расширенному и дополненному в расчете на транскрибирование связного текста.

С учетом упомянутых выше требований к тексту - ограниче - ния на длину фразы, простота синтаксиса, удобопроизносимость

слов - первоначально был вручную (как начальное приближение) сформирован массив, содержащий около 400 фраз. Статистический анализ текста показал, что одни биграммы (как правило, внутрисловные) встретились по 300 и более раз каждая; другие сочетания (на границах слов) встретились по 1-2 раза, а 41% биграмм не встретились вообще.

Прояснилась также динамика накопления редких биграмм. Так, в тексте объемом 100 фаз они составляют 20%, 200 - 18%, 300 - 15%, 400 - 10%. Таким образом, дальнейшее наращивание объема текста дает неизбежное увеличение числа частых сочетаний и все более незначительный прирост встречаемости "нужных" (ранее не встретившихся) биграмм.

Эта непреложная зависимость обусловила поиск иного подхода к формированию текста со стопроцентным содержанием всех биграмм. Суть его заключалась в направленном составлении фраз, включающих как можно больше редких сочетаний фонем. (Их перечень был известен из предыдущих статистических работ авторов с текстами и со словарем русского языка Д.Уорта [4,6].) Частые биграммы, "обрамляя" в словах и словосочетаниях редкие, будут накапливаться сами собой.

Словарь Д.Уорта, записанный поморфемно в орфографическом виде и в транскрипции, содержится на магнитных носителях в ИМ СО АН СССР и НГУ. Посредством считывания из словаря Д.Уорта слов, содержащих редкие биграммы, вручную формировались содержащие их фразы, а после формирования всего текста с помощью машинной обработки в нем автоматически выделялись слова и целые фразы, не содержащие "нужных" биграмм. Затем (вручную) эти "пустые" участки опускались, если это не нарушало смысла и грамматики фразы, или заменялись другими, содержащими не встречавшиеся биграммы.

Следует отметить, что внесение даже незначительных поправок влекло за собой изменения в уже сложившейся статистической

картине текста, и для получения окончательных результатов - максимально сжатого текста со стопроцентной встречаемостью всех возможных в речи двухфонемных сочетаний - потребовалось множество ручных и машинных итераций.

Итак, получен текст со следующими характеристиками:

- время звучания - около 20 минут;
- число фраз - 357;
- слов во фразе в среднем 5;
- предложения простые распространенные;
- общее число биграмм - 11838, в том числе с частотой встречаемости: 100-10 - 18%, 9-2 - 41%, единичных - 41%.

Таким образом, 2013 возможных в речи *разных* двухфонемных сочетаний реализовано в 357 фразах, содержащих около 12 тысяч биграмм.

Приведем примеры фраз из подготовленного текста:

№ 28. Сфигмограф выдал ленту.

№ 114. В оценках и характеристиках Георгия - пейоративный оттенок.

№ 272. Семь Симеонов не сходили с языка, но потом забылись.

№ 295. Печь зияла черным зевом.

Заключение

В заключение следует отметить, что значение построения универсального текста шире, чем указано в цели работы. Фонограмма текста может использоваться (и уже использовалась) для подготовки эталонов в системе микроволнового синтеза, так как она имеет все необходимые стационарные и переходные участки в акустическом сигнале русской речи. Текст может использоваться не только в системах распознавания, основанных на модели скрытых марковских процессов, но и в традиционных системах распознавания типа "анализ через синтез" для автоматического компилирования

эталонов возможных слов и фраз слитной речи по контексту распознаваемого диалога. Как и в модели скрытых марковских процессов, эти приложения требуют членения (желательно, автоматизированного) фонограммы текста на фонемы, что составляет отдельную, довольно трудоемкую, задачу, которую мы здесь не рассматриваем.

Направление дальнейших работ предполагает решение поставленной задачи в полном объеме, т.е. с набором представительной выборки биграмм не менее 5 для самых редких при минимальном объеме текста. Попробуем оценить необходимый объем текста, исходя из полученных в данной работе результатов.

Сравнение с английским аналогом [3] требует внесения следующей коррекции. Инвентарь фонем в англоязычных системах распознавания содержит 41 фонему (плюс пауза #). Если соблюдать пропорции, то русский текст для 56 фонем должен быть в $(56/42)^2 = 1{,}78$ раз длиннее, т.е. составлять около 35 минут. Минимальная длина текста для 5-кратного повторения 2000 биграмм (или фонем, что по длительности одинаково) - около 1000 секунд, или 17 минут. Вероятно, достаточно взять коэффициент 2 на часто встречающиеся биграммы, что приведет к той же грубой оценке. С эргономической точки зрения двукратное увеличение длины текста по сравнению с априорно заданной существенно снижает характеристики обучающей системы. Поэтому вслед за отказом от требования легкой произносимости, возможно, придется отказаться и от средней длины фраз в 5 слов в сторону ее уменьшения.

Возможности подготовленного текста с точки зрения полноты выполнения требований выглядят следующим образом. 807 единичных биграмм внутри слов встречаются 233 раза (29%), на границах слов - 574 раза (71%). Все слова, фразы, словосочетания, содержащие единичные биграммы, составляют 59% от всего объема текста (7031 биграмма). Время звучания около 12 минут. Таким образом, если механически повторить этот текст 5 раз, то требуе-

мая представительность выборки будет получена. При этом общая длительность текста будет около 70 минут, что в 2 раза превышает ориентировочную оценку. Не прекращая усилий по уменьшению объема универсального текста, мы намереваемся оценить качество полученного текста в реальных приложениях с использованием марковской модели. Вряд ли все возможные в реальном сигнале последовательности наблюдаемых величин будут отражены даже в полном представительном озвученном тексте, содержащем все последовательности скрытых состояний. Обычный прием в случае,когда при распознавании встречается не представленная в обучающей фонограмме последовательность наблюдаемых величин, состоит в приписывании этой последовательности малого, но ненулевого,заранее оцененного значения вероятности [7]. Не исключено, что применение этого приема существенно снизит требования к обучающему тексту, не слишком ухудшив его качество. Ответ будет получен экспериментально.

Литература

1. ВЕЛИЧКО В.М., САЛОМАТИНА Н.В., ЮДИНА Л.С. Автоматиза - ция выбора обучающей последовательности при распознавании слитной речи //Автоматическое распознавание слуховых образов. Тезисы докладов 15-го Всесоюзного семинара (АРСО-15).Таллинн,1989, - С. 265-266.

2. ВЕЛИЧКО В.М., ЗАГОРУЙКО Н.Г., ФОСС С.Г. Статистический подход к распознаванию речи //Там же. - С. 13-17.

3. CUBALA F., et.al. Continious Speesh Recognition Results of the BYBLOS System on the DARPA 1000-Word Resource Management Database //IEEE. ICASSP-88. Ser. 7.8.

4. ЁЛКИНА В.Н., ЮДИНА Л.С., ХАЙРЕТДИНОВА А.Г. Статистика двух- и трехфонемных сочетаний русской речи //Вычислительные системы. - Новосибирск, 1969. - Вып. 37. -С. 48-53.

5. ЮДИНА Л.С. Автоматизация транскрипции и некоторые количественные характеристики русской речи: Автореф. дисс. Филолог.н. 10.02.01. - Новосибирск, 1973. - 23 с.

6. WORTH D., KOZAK A., JOHNSON D. Russian Derivation Di - ctionary. - New-York, 1970. -747 p.

7. ДЖЕЛИНЕК Ф., (Ф.ЕЛИНЕК). Распознавание непрерывной речи статистическими методами //ТИИЭР. - 1976. - Том 64, № 4. - С. 131-160.

Поступила в ред.-изд.отд.
18 июня 1991 года