Применение синтаксического анализа в задаче распознавания речи

Зулкарнеев М.Ю., *кандидат физико-математических наук.*

Репалов С.А., *кандидат физико-математических наук.*

Шамраев Н.Г., *аспирант.*

Статья посвящена методам повышения точности распознавания речи, основанным на применении синтаксического анализа предложения. Рассмотрены модификации классической N-граммной языковой модели с помощью синтаксической модели предложения, которая строится на основе вероятностных грамматик Хомского. Рассмотрены различные варианты применения алгоритма Коке-Янгера-Касами для синтаксического разбора предложения. Показано преимущество построенной языковой модели перед классической N-граммной моделью.

• распознавание речи • синтаксический анализ • алгоритм Коке-Янгера-Касами.

The article is devoted to methods of speech recognition accuracy improvements, based on the use of parsing sentences. Modification of the classical N — gram language model using syntactic model, based on probabilistic Chomsky grammar, is presented. The different versions of the Cocke-Younger-Kasami algorithm application for sentence parsing are discussed. The advantage of the proposed language model over classical N-gram model is shown.

• speech recognition • syntacticol analysis • Coche-Younger-Kasami algorithm.

Введение

В настоящее время в распознавании речи широко распространён метод, использующий скрытые марковские модели и n-граммные модели языка [1]. Хотя использование этого метода позволило достичь определённых успехов при создании систем распознавания речи, используемые модели обладают многими недостатками. В частности, n-граммная модель языка способна описывать зависимости между словами, расстояние между которыми не превышает некоторого числа.

В данной работе исследуется возможность моделирования языка при помощи синтаксических моделей, представленных в виде вероятностных контекстно-свободных грамматик Хомского (PCFG — probabilistic context-free grammar) и, в частности, использование этих моделей в задаче распознавания речи. Для этого предлагаются и исследуются два метода, позволяющих использовать синтаксическую модель языка, представленную в виде PCFG, в задаче распознавания речи, а также приводятся результаты экспериментов, показывающих преимущество предлагаемого подхода по сравнению с подходом, основанным на n-граммных моделях языка.

Описание алгоритма Коке-Янгера-Касами

Классический алгоритм Коке-Янгера-Касами (CKY) предназначен для обработки линейных последовательностей слов. Впервые он приводится в работе Янгера [2]. В статье мы рассмотрим реализацию алгоритма для контекстно-свободных грамматик в нормальной форме, т.е. когда каждое правило имеет либо вид $A \to BC$, либо $A \to \beta$, где A, B, C — нетерминальные символы, β — терминальный символ.

Задача рассматриваемого алгоритма состоит в ответе на вопрос: может ли данное предложение W быть сгенерировано грамматикой Хомского, заданной фиксированным набором таких правил.

Пусть $\Psi_{ij}(A)$ обозначает некоторую строку α, генерируемую правилом A, начинающуюся с ??-го слова и заканчивающуюся на *j*-м слове в исходной последовательности слов W. $\varphi_{ij}(A)$ — функция стоимости вывода этой последовательности правилом A. Для задачи синтаксического анализа предложения в качестве функции стоимости выступает логарифм вероятности правила перехода нетерминальных классов в терминалы (слова).

В классическом алгоритме начальные значения функций устанавливаются для каждого слова по следующим формулам:

$$\varphi_{ii}(A) = \min_{\{A \to \alpha\}} \{C[\alpha | t_{i-1,i}]\},$$

$$\psi_{ii}(A) = \alpha = \arg\min_{\{A \to \alpha\}} \{\varphi_{ii}(A)\}, \text{ для } 1 \le i \le |W|.$$

Все остальные значения Ψ и φ не определены в этот момент. Далее уже для пар $1 \le i < j \le |W|$ последовательно вычисляются

$$\varphi_{ij}(A) = \min_{\{A \to BC\}} \{\min_{i \le l < j} \{\varphi_{il}(B) \cdot \varphi_{l+1,j}(C)\}\}, \quad (1)$$

$$\psi_{ij}(A) = \psi_{il}(B) \otimes \psi_{l+1,j}(C). \quad (2)$$

Т.е. идёт перебор по l-индексу слова-разделителя и всевозможным правилам B и C, таким, что $A \to BC$. Вычисление начинается с пар, таких, что $|i-j| = 1$, затем для $|i-j| = 2$ и так далее по возрастанию до $|i-j| = |W| - 1$.

Если правило A генерирует подстроку α_{ij}, то для него найдётся тройка оптимальных объектов $(\bar{l}, \bar{B}, \bar{C}) = \arg\min_{\substack{\{A \to BC\}, \\ i \le l < j}} \{\varphi_{ij}(A)\}$ — индекс слова разделителя l, и два правила B и C. Если предложение W может быть сгенерировано из стартового символа S, т.е. $S \Rightarrow W$, то алгоритм восстановления бинарного дерева будет следующим.

Рассмотрим тройку для стартового слова $(l, B, C) = \psi_{1N}(S)$, где $N = |W|d$. Используя это разбиение, перейдём к левому поддереву $(l_1, B_1, C_1) = \psi_{1l}(B)$ и к правому поддереву $(l_2, B_2, C_2) = \psi_{l+1N}(C)d$.

Далее рекурсивно разбираем каждую тройку, пока не получим последовательность терминальных символов.

Подход на основе линейной последовательности слов

Первый подход к использованию синтаксического анализа для распознавания речи заключается в использовании алгоритма CKY для линейной последовательности псевдослов, которые получаются на основе решёток слов, генерируемых декодером.

Результатом работы декодера является решётка слов, при этом каждому слову сопоставлены два значения — акустическая и лингвистическая вероятности. Для оценки лингвистической вероятности при построении решётки используется 3-граммная модель языка. Пример решётки приведён на рис.1.

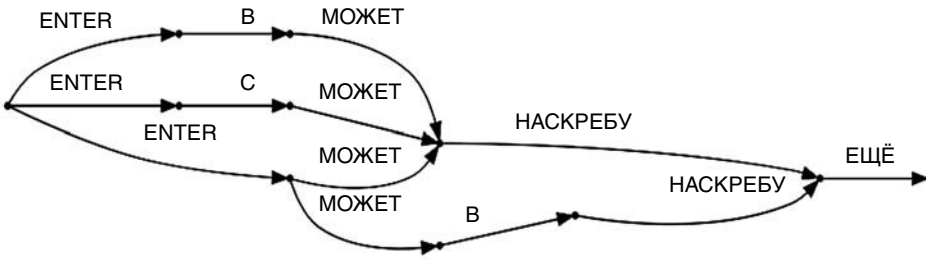

Рис. 1. Пример решётки слов

При помощи алгоритма кластеризации [3] эта решётка преобразуется в линейную сеть (Word Confusion Network, WCN), которая является компактным представлением множества гипотез. Представление гипотез в виде всех возможных предложений и последующий их последовательный анализ не эффективен с точки зрения быстродействия. Если множество гипотез будет слишком маленьким, снижается вероятность найти правильную гипотезу, если множество гипотез слишком большое, очень сильно (экспоненциально) возрастает время, затрачиваемое на синтаксический разбор.

Один из предлагаемых здесь способов ускорения работы и прямого использования метода CKY состоит в возможности представления множества гипотез в виде линейной сети слов. В такой сети слова разбиты на группы, которые упорядочены во времени. Пример такой сети приведён на рис. 2. На нём цифрами обозначены вероятности, символом eps — пропуск слова.

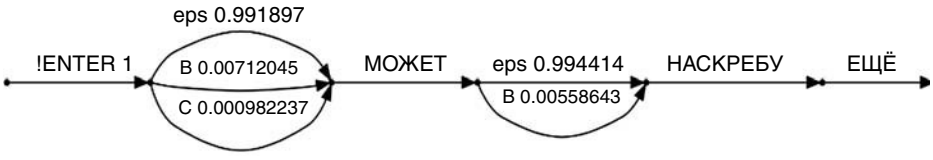

Рис. 2. Пример линейной решётки слов

Предварительная подготовка для обработки гипотез заключается в определении множества правил, записанных в форме контекстно-свободной грамматики Хомского для конкретного языка и вычисление вероятностей всех правил,

нетерминальных и терминальных, входящих в эту грамматику (осуществляется с помощью метода Бейкера для большого массива предложений).

На этапе анализа линейная решётка слов (см. рис. 2) преобразуется в последовательность линейно присоединённых элементов. Каждый элемент соответствует некоторому фрагменту предложения и содержит набор конкурирующих слов. Это позволяет представить решётку как линейную последовательность некоторых формальных символов:

$$W_1, W_2, \ldots, W_N. \quad (3)$$

Например, формальное слово W_1 может включать в себя слова $word_1$, $word_2$, …, $word_{K1}$. Далее все возможные слова, соответствующие W_i, будут генерироваться некоторым множеством предтерминальных классов. Тогда в общее множество правил добавляются порождающие терминальные правила вида $S_{ij} \rightarrow W_1$. Вероятность каждого правила вычисляется по следующей формуле:

$$\sum_{word_i \in W_1} P_{dic}^{(i)} P_{lat}^{(i)},$$

где $word_i$ — некоторое слово, находящееся в последовательности на первом месте, j — индекс перебора множества предтерминальных классов, генерирующих все слова $word_i$, S_{i1} — соответствующий грамматический (предтерминальный) класс для $word_1$, $P_{dic}^{(i)}$ — вероятность $word_i$ в словаре, $P_{lat}^{(i)}$ — вероятность $word_i$ в линейной решётке. Так как одно и то же слово может генерироваться разными грамматическими классами, то нужно рассматривать все допустимые варианты.

После пополнения таким способом множества правил, для последовательности символов (3) её синтаксический разбор становится возможен. Он осуществляется при помощи вероятностного алгоритма Коке-Янгера-Касами. На выходе получается наиболее вероятное дерево правил, генерирующее последовательность (3). Это дерево представляет собой фактически синтаксический разбор предложения. Таким образом, мы получаем наиболее вероятную последовательность грамматических классов $G_1, G_2, \ldots, G_N$ для исходного предложения. Далее, для восстановления оптимальной гипотезы из слов-кандидатов, принадлежащих формальному слову W_i, выбирается слово, обладающее максимальной вероятностью среди всех слов, принадлежащих грамматическому классу G_i.

К сожалению, данный метод синтаксического анализа имеет серьёзные недостатки:

1. Преобразование исходной решётки в линейную последовательность слов приводит к потере части гипотез.
2. Вторым следствием преобразования является фиксированная длина гипотезы предложения. Определённая гибкость сохраняется лишь при переходе псевдослова в «пустой» класс eps, т.е. только для уменьшения фиксированной длины.
3. Фиксация грамматических классов на последнем этапе может привести к гарантированно неоптимальному выбору в случае, если истинное слово принадлежит другому классу.

Недостатки данного метода можно преодолеть, если использовать более гибкий подход, сохраняющий на каждом этапе большую информацию о конкурентных гипотезах. В качестве такого подхода мы предлагаем модификацию метода CKY для решётки гипотез, рассматриваемую в следующем разделе.

Синтаксический разбор решётки слов

Классический метод Коке-Янгера-Касами предназначен для обработки линейных последовательностей слов. В таком виде он не очень удобен для применения к задаче распознавания речи. Метод прямого перебора гипотез оказывается неэффективным с точки

зрения времени обработки [4]. Метод, основанный на WCN, позволяет эффективно повысить скорость обработки. Однако он является искусственным и обладает рядом недостатков, самый важный из которых — фиксация цепочки грамматических классов и подбор гипотезы в зависимости от этой цепочки. Это делает метод сильно зависящим от ошибок синтаксического разбора.

В работе [3] предлагается обобщение алгоритма Коке-Касами-Янгера, в котором в качестве анализируемых сегментов выступают временные интервалы, а не цепочки слов. Эта модификация формально не меняет расчётные формулы (1), (2), а приводит только к увеличению размеров матрицы $\Psi_{ij}(A)$ (теперь она будет равна T x T, где T — количество дискретных моментов времени).

В данной работе на анализируемые временные интервалы накладывается ограничение в виде решёток слов, которые задают возможные цепочки слов для данного временного интервала. Данному временному интервалу соответствует цепочка слов, если в решётке имеются узлы, соответствующие началу и концу временного интервала, и они связаны переходами. Ограничение приводит к тому, что размер матрицы $\Psi_{ij}(A)$ будет равен количеству узлов в решётке. Результатом работы алгоритма Коке-Касами-Янгера, в этом случае, будет наиболее вероятный, с точки зрения PCFG, путь в решётке. По аналогии с синтаксическим разбором предложения здесь можно говорить о синтаксическом разборе решётки слов.

Рассмотрим работу описанного алгоритма Коке-Касами-Янгера на решётках на примере разбора решётки слов, представленной на рис. 3. Узлы решётки обозначены прямоугольниками и соответствуют времени конца слов, которыми они подписаны. Переходы между узлами фактически соответствуют словам в речевом сообщении. Узлы решётки пронумерованы в соответствии с топологией решётки (номер начального узла любого перехода меньше, чем номер конечного узла этого перехода).

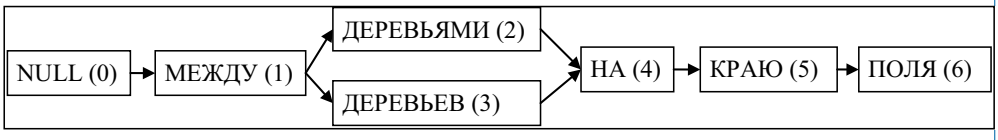

Рис. 3. Решётка слов

В алгоритме используется матрица D, которая хранит информацию о том, какие грамматические классы и с какой вероятностью соответствуют различным частичным путям в решётке. Например, элемент матрицы $D[0,2]$ хранит список грамматических классов, которые могли бы соответствовать цепочке слов «МЕЖДУ ДЕРЕВЬЯМИ».

Работа алгоритма начинается с инициализации элементов матрицы D, которые соответствуют переходам между узлами. В них заносится информация о возможных грамматических классах, которым могут принадлежать слова переходов. Например, элементу матрицы $D[0,1]$ соответствует переход $(0,1)$ (слово «МЕЖДУ»). В этот элемент заносится класс PR (предлог) и класс ADV (наречие) с соответствующими вероятностями.

Далее идёт основная часть алгоритма, которая состоит в следующем. Рассматриваются всевозможные пары узлов по возрастанию разности их номеров, начиная с разности, равной 2, и для каждой такой пары (i,j) рассматриваются всевозможные узлы k, которые могут лежать на частичных путях, соединя-

ющих эти узлы. Для этой пары узлов (i,j) принимается решение: может ли ей соответствовать некоторый грамматический класс, исходя из имеющихся правил и информации, хранящейся в элементах матрицы $D[i,k]$ и $D[k,j]$. Например, при рассмотрении пары узлов $(0,2)$ принимается решение, что цепочке слов «МЕЖДУ ДЕРЕВЬЯМИ» соответствует грамматический класс *abst* (обстоятельство), поскольку существует правило *obst* → *PR predl_mn_sred_tvor_neod*, а в элементах $D[0,1]$ и $D[1,2]$ хранится информация о грамматических классах *PR* (предлог) и *predl_mn_sred_tvor_neod* (неодушевлённая предложная фраза множественного числа, среднего рода, творительного падежа) соответственно. Описанная процедура продолжается до тех пор, пока не будет рассмотрена пара самых удалённых узлов $(0,6)$.

На этом синтаксический разбор решётки завершён, и дальше информация, сохранённая в матрице D, используется для восстановления наиболее вероятной цепочки слов. Также при необходимости информация из D может быть использована для восстановления наиболее вероятного дерева синтаксического разбора.

Полученные результаты

В данной работе алгоритм Коке-Касами-Янгера на решётках слов включён в процесс распознавания, как постобработка решёток слов, которые получены на первом этапе распознавания, направленная на уточнение результатов распознавания. В качестве функции $\Psi_{ij}(A)$ берётся комбинация языковой вероятности, полученной при помощи трёхграммной модели языка и вероятности синтаксических правил. Это позволяет объединять информацию, заложенную в трёхграммной модели и в синтаксической модели и, таким образом, получать более надёжные результаты.

Для проверки предложенных подходов были проведены эксперименты по распознаванию программ на русском языке, записанных из прямого эфира радио «Свобода». В качестве базовой была взята система распознавания русской речи, использующая связанные трифоны и трёхграммную модель языка. Синтаксическая модель русского языка, представленная в виде PCFG, была сгенерирована автоматически на основе синтаксически размеченных предложений Национального корпуса русского языка [5, 6].

Эксперименты показали, что применение метода синтаксического разбора, основанного на превдословах, позволило повысить точность с 72,39 до 75,16%. В основном, повышение точности было достигнуто за счёт согласования падежей прилагательных и существительных, употребления предлогов.

Для проверки подхода, основанного на синтаксическом разборе решёток слов, была создана программная реализация метода Коке-Касами-Янгера синтаксического разбора решёток слов.

Из-за сложности алгоритма Коке-Янгера-Касами, который составляет $O(N^3)$, где N — количество узлов в решётке, реализация не позволила провести полноценные эксперименты для больших решёток. Однако эксперименты для усечённых решёток показали преимущества подхода по сравнению с n-граммными моделями языка. Подход позволяет устанавливать зависимость между значительно удалёнными друг от друга словами в предложении.

Заключение

Использование синтаксического анализа при распознавании речи даёт дополнительный источник информации, который позволяет значительно повысить точность распознавания речи. Обобщение метода Коке-Янгера-Касами на решётки слов — наиболее удобный механизм включения синтаксических правил в систему распознавания речи. Однако его сложность не позволила выполнять анализ полноценных решёток слов.

В качестве дальнейшего развития планируется расширить алгоритм, включив в него дополнительные механизмы отсечки маловероятных гипотез, которые бы значительно ускорили обработку, но при этом не привели к потерям правильных гипотез.

Литература

1. *Rabiner L.R., Juang B.H*. Fundamentals of Speech Recognition. Prentice-Hall, Inc. Upper Saddle River, NJ, USA, 1993.

2. *Younger D*. Recognition and parsing of context free languages in time n3. Information and Control. Volume 10. Issue 2. February 1967. Pp. 189–208.

3. *Levenson S.C*. Mathematical models for speech technology. John Wiley & Sons Ltd, NJ, USA, 2005.

4. *Батальщиков А.А., Зулкарнеев М.Ю., Шамраев Н.Г*. Оценка гипотез с использованием синтаксического анализа // Сборник трудов XXII сессии Российского акустического общества и Сессии Научного совета РАН по акустике. Т. 3. М.: ГЕОС, 2010. С. 22–25.

5. Национальный корпус русского языка: 2006–2008. Новые результаты и перспективы. СПб.: Нестор-История, 2009. 502 с.

6. http://www.ruscorpora.ru/

Сведения об авторах

Зулкарнеев Михаил Юрьевич —
старший научный сотрудник ФГНУ НИИ «Спецвузавтоматика», г. Ростов-на-Дону. Область научных интересов — распознавание и анализ устной речи.

Репалов Сергей Анатольевич —
заведующий лабораторией ФГНУ НИИ «Спецвузавтоматика», г. Ростов-на-Дону. Область научных интересов — автоматическая и инструментальная обработка устной речи, цифровая обработка сигналов.

Шамраев Николай Георгиевич —
научный сотрудник ФГНУ «НИИ «Спецвузавтоматика», г. Ростов-на-Дону. Область научных интересов — распознавание и анализ устной речи.