Вопросы для kmp

Здравствуйте, Михаил Петрович!

Меня интересуют два вопроса:

1.какое количество слов может распознать компьютерная программа по распознаванию и синтезу речи?

2 существуют ли грамматические правила для компьютерных программ распознания речи и какие они (пример)?

kmp-ответ №1

Компьютерные программы и синтеза речи, и распознавания речи могут распознать разное количество слов: от 0 (если они ….) до бесконечности (если они используют рекурсивные правила).

Но какое количество слов нам нужно распознавать?

Объем лексикона

- полтора года - 100 слов
- два года - 300-400
- три года - 1000-1100
- четыре года - 1600
- пять лет – 2200
- выпускник средней школы - 3500-5000 слов.
- человека с высшим образованием - 8000-12000 слов

Объем лексикона

- в "Словаре языка А.С. Пушкина" в 4-х томах (М., 1956-1961) - 21 191 слово.
- в произведениях Вильяма Шекспира - 29 066 лексем.
- в произведениях Иоганна Вольфганга Гёте - 17 000 слов.
- словарь Эллочки-людоедки составлял лишь 30 слов, но ими она могла выразить практически любую свою мысль.

Словарная лексика

- Словарь современного русского литературного языка (Большой академический словарь, БАС) в 17 томах (с 1948 по 1965) 131 257.
- «Толковый словарь живого великорусского языка» В. И. Даля 200 000 слов.
- "Частотный словарь русского языка" под ред. Л. Н. Засориной - 30 000 слов (6 000 покрывают 90% обработанных).

Wikipedia

На 1 ноября 2013 года:

- Английский раздел: 4 334 000 статей
- Немецкий раздел: 1 633 000 статей
- Русский раздел: 1 046 000 статей

MultiTran

8 000 000 слов

800 тематик (прикладных областей), в том числе:

- Авиационная медицина 28 559 слов
- Вычислительная техника 105 782слов
- Лингвистика 10 279 слов
- Строительство 100 801 слов

Доменная номинация

- IP-адрес: уникальное число, однозначно идентифицирующее компьютер в Интернете.
- В IPv4 возможно 2^{32} адресов = 4,3 млрд.
- Они давно исчерпаны и смягчает проблему нехватки IP-адресов технология NAT (Network address translation).
- 04.02.2008 начат переход к IPv6, который обеспечит 2^{128} = 340 триллионов триллионов триллионов.
- Доменные имена: однозначно соответствующие IP-адресу имена компьютеров в Интернете.

Терминологическое...

Термины

Номены

Прототермины

Терминоиды

Предтермины

Квазитермины

Онимы

Хрематонимы

Вывод

- Светлое познается светлым (Плотин)
- Живое ….. живым
- Бесконечное …… бесконечным

- Тёмным - тёмное
- Мёртвым – мёртвое
- Конечным – конечное (уж Концевой-то знает)

Второй вопрос

2 существуют ли грамматические правила для компьютерных программ распознания речи и какие они (пример)?

kmp-ответ №2

Для программ правил не существует, правила могут использоваться в программах.

Это:

- правила формальных грамматик,
- рекурсивные правила,
- алгоритмы скрытых марковских моделей.

Грамматика

Грамматический строй
(**грамматическая система**, **грамматика**;
от греч. γράμμα — запись)

- совокупность языковых средств языка, регулирующих правильность построения значимых речевых отрезков (слов, высказываний, текстов).

Формальные грамматики

Формальная грамматика — способ выделения некоторого подмножества из множества всех слов некоторого конечного алфавита:

- порождающие грамматики — задают правила, с помощью которых можно построить любое слово языка
- распознающие (аналитические) грамматики — позволяют по данному слову определить, входит оно в язык или нет.

Иерархия Хомского (1959)

Грамматики делятся на 4 типа, каждый последующий является более ограниченным подмножеством:

- тип 0. неограниченные грамматики — возможны любые правила
- тип 1. контекстно-зависимые грамматики.
- тип 2. контекстно-свободные грамматики.
- тип 3. регулярные грамматики — самые простые (эквивалентны конечным автоматам).

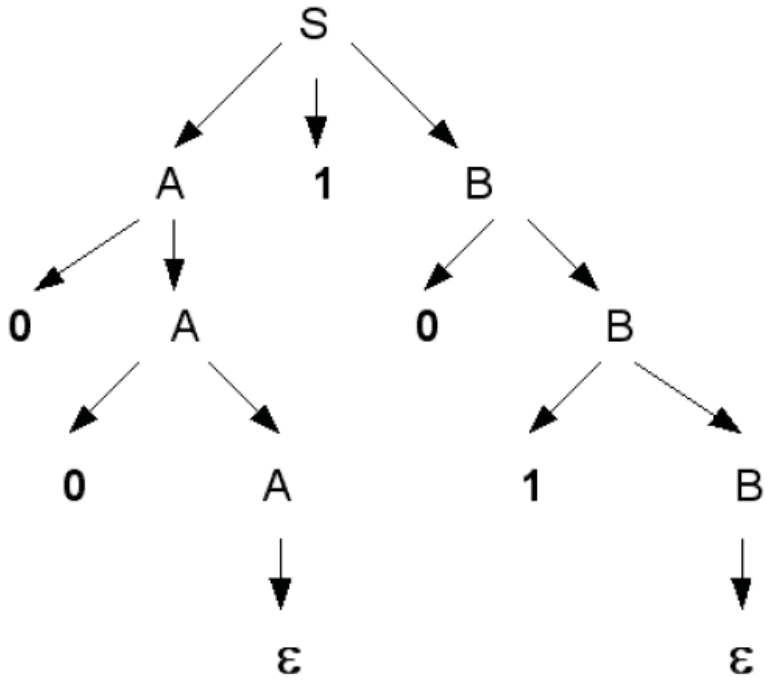
Контекстуальное…

- Контекстно-свободные грамматики применяются для определения грамматической структуры в грамматическом анализе.
- Регулярные грамматики применяются для текстового поиска, разбивки и подстановки, в том числе в лексическом анализе.

Рекурсивное

- Основным отличием человеческого языка от языков иных видов является открытость и рекурсивность.
- Рекурсивно перечислимый язык (RE, частично разрешимый). В иерархии Хомского - язык типа 0.
- Рекурсивный язык (RP, не определен в ИХ) — для которого существует машина Тьюринга, останавливающаяся на любой входной цепочке, когда она принадлежит языку.
- Все регулярные, контекстно-свободные и контекстно-зависимые языки рекурсивны.

Пример №1



Пример №2

$$L_1 = \{ab, bc, cd, de\}$$
$$L_2 = \{a^i b^i c^i d^i e^i \mid i > 0\}$$
$$L_3 = \{\omega\omega^R \mid \omega \in \{a, b, c, d, e\}^*\}$$
$$L_4 = \{(ae)^i (bdc)^j \mid j - 1 = i \geqslant 0\}$$
$$L_5 = \{dead, bad\}$$

Пример №3

$$
\begin{aligned}
&S \to 0A' \mid 0D' \mid \ldots \mid 9D' \\
&A' \to xB' \mid 0C' \mid \ldots \mid 7C' \\
&B' \to 0B' \mid 1B' \mid \ldots \mid 9B' \mid aB' \mid \ldots \mid fB' \mid AB' \mid \\
&0z' \mid 1z' \mid \ldots \mid 9z' \mid az' \mid \ldots \mid fz' \mid Az' \mid \ldots \mid Fz' \\
&C' \to OC' \mid \ldots \mid 7C' \mid z' \\
&D' \to OD' \mid \ldots \mid 9D' \mid z' \\
&z' \to u \mid l \mid ul \mid lu \mid U \mid L \mid UL \mid LU \mid \varepsilon
\end{aligned}
$$

Марков Андрей Андреевич

Русский математик, академик.

Годы жизни: 1856-1922

Первооткрыватель стохастических процессов:

- следующее состояние процесса зависит, вероятностно, только от текущего состояния.

Скрытая марковская модель

- **СММ** — статистическая модель, имитирующая работу процесса, похожего на Марковский процесс с неизвестными параметрами, и задачей ставится разгадывание неизвестных параметров на основе наблюдаемых.
- СММ применяются в области распознавания речи, письма, движений, машинном переводе, биоинформатике, криптоанализе.

Скрытая марковская модель

Скрытая модель Маркова — это вероятностная модель множества случайных переменных $\{O_1, \ldots, O_t, Q_1, \ldots, Q_t\}$. Переменные O_t — известные дискретные наблюдения, а Q_t — «скрытые» дискретные величины. В рамках скрытой модели Маркова есть два независимых утверждения, обеспечивающих сходимость данного алгоритма:

1. t-я скрытая переменная при известной $(t-1)$-ой переменной независима от всех предыдущих $(t-1)$ переменных, то есть
$$P(Q_t \mid Q_{t-1}, O_{t-1}, \ldots, Q_1, O_1) = P(Q_t \mid Q_{t-1})$$
2. t-е известное наблюдение зависит только от t-го состояния, то есть не зависит от времени, $P(O_t \mid Q_t, Q_{t-1}, O_{t-1}, \ldots, Q_1, O_1) = P(O_t \mid Q_t)$

Примеры алгоритмов СММ

- Алгоритм Витерби: даны параметры модели, требуется определить наиболее подходящую последовательность скрытых узлов, наиболее точно описывающую данную модель (помогает при решении данной задачи).
- Алгоритм Баума-Велша: дана выходная последовательность (или несколько) с дискретными значениями, требуется «потренировать» СММ на данном выходе.