both are either vacant or have a disabled holder of that office, the next officer in the Presidential line of succession, the Speaker of the House, becomes Acting President. The line extends to the President pro tempore of the Senate after the Speaker, followed by every member of the Cabinet in a set order.

References

1. *Козикис Д. Д.* Страноведение США = American Studies: учебное пособие/ Д.Д.Козикис, Г.И.Медведев, Н.В.Демченко. – Минск.: Лексис, 2008. – 272 с.
2. Интернет-адрес: http://en.wikipedia.org/wiki/President_of_the_United_States.
3. United States Information Agency**,** An Outline of American Government/ edited by Richard C. Schroeder. Revised and updated in 1989 by Nathan Glick.

СТАТИСТИЧЕСКАЯ МОДЕЛЬ ЯЗЫКА

Я. Г. Кунцевич, Д. А. Кушнарёв

ВВЕДЕНИЕ

В компьютерной лингвистике одной из самых важных задач является задача о построении компьютерных моделей языка. Модели бывают разных видов, и выбор одной из них зависит от решаемой проблемы.

В данной работе будет кратко рассказано о различных вариантах статистической модели языка, которая позволяет определить вероятность появления в тексте последовательности языковых единиц.

С полной версией работы можно ознакомиться по адресу [1].

ЧТО ТАКОЕ СТАТИСТИЧЕСКАЯ МОДЕЛЬ ЯЗЫКА

Статистическая модель языка определяет вероятность последовательности m слов P(w1,…,wn) посредством распределения вероятности или иными словами распределение вероятности различных последовательностей слов.

В качестве модели языка в системах статистического перевода используются преимущественно n-граммные модели, утверждающие, что грамматичность выбора очередного слова при формировании текста определяется только тем, какие (n – 1) слов идут перед ним. Вероятность каждого n-грамма определяется по его встречаемости в тренировочном корпусе [2].

КРИТЕРИИ ОЦЕНКИ КАЧЕСТВА МОДЕЛЕЙ

Модель можно оценить по следующим критериям:

1. Качество модели – кросс-энтропия на тестовом тексте.
2. Полнота модели (число слов, которые модель знает).
3. Число ошибок в прикладных программах.

N-ГРАММНЫЕ МОДЕЛИ

В модели n-грамма вероятность наблюдения предложения w1,…,wn приближена к:

$$P(w_1,\ldots,w_n) = \prod P(w_i \mid w_1,\ldots,w_2) \approx \prod P(w_i \mid w_{i-(n-1)}, \ldots, w_{i-1}).$$

Условная вероятность может быть вычислена от счета частоты n-грамма:

$$P(wi \mid wi-(n-1) ,\ldots, wi-1) = count\ (wi-(n-1), wi-1, \ldots,wi)/count\ (wi-(n-1), \ldots, wi-1).$$

Слова «биграмм и триграмм языковая модель» обозначают языковые модели n-грамма с n=2 и n=3, соответственно.

Вероятность предложения «*I like to play chess*» в триграммной (n=3) языковой модели рассчитывается так:

$$P(I, like, to, play, chess) \approx$$
$$P(I)*P(like|I)*P(to|I, like)*P(play|like, to)*P(chess| to, play).$$

Языковые модели, основанные на n-граммах, используют явное предположение о том, что вероятность появления очередного слова в предложении зависит только от предыдущих n-1 слов. На практике используются модели со значениями n = 1, 2, 3 и 4. Наиболее удачной моделью из этого класса оказывается *триграммная модель*. На сегодняшний день практически все коммерческие системы распознавания речи используют n-граммную модель в той или иной форме.

Основным достоинством данного класса моделей оказывается возможность построения модели по обучающему корпусу достаточно большого размера и высокая скорость работы.

Основные недостатки – заведомо неверное предположение о независимости вероятности очередного слова от более длинной истории, что не позволяет моделировать глубокие связи в тексте и большие объёмы данных. Для уменьшения размеров моделей используют техники сглаживания, которые позволяют производить оценку параметров модели в условиях недостаточных данных. Другой подход – кластеризация словаря [3].

ДАЛЬНОДЕЙСТВУЮЩАЯ ТРИГРАММНАЯ МОДЕЛЬ

Эта вероятностная модель языка расширяет триграммные модели, предсказывая слова не только по двум непосредственно предшествующим словам в предложении, но и потенциально по любой паре стоящих рядом слов, которые лежат внутри этого же предложения.

Грамматика расширяет триграммные модели через разрешение связей между словами, потенциально находящимися на большем расстоянии от предсказываемого слова в пределах предложения. Дальнодействующая

триграммная модель может пропускать малоинформативные слова и улучшать предсказуемость в модели.

МОДЕЛИ ДРУГИХ УРОВНЕЙ ЯЗЫКА

Описанные выше модели определяли вероятность появления предложения или словосочетания, но можно построить и модели языка на других его уровнях. Возможны следующие варианты:

1. Модель, которая позволяет определить вероятность появления слова в данном языке на основе анализа частоты и порядка появления символов (графем) или фонем в данном языке.

Например, с помощью такой модели можно определить вероятность появления в русском языке слова «лингвистика»: P(лингвистика) = P(л)*P(л,и)*P(л,и,н)*P(и,н,г)* P(н,г,в)* P(г,в,и)* P(в,и,с)* P(и,с,т)* P(с,т,и)* P(т,и,к)* P(и,к,а) (в случае использования триграммной модели).

Главным недостатком такого вида моделей является их низкая точность. Преимуществами – компактность и универсальность: программу для работы с таким видом моделей легко переориентировать на работу с другими языками.

2. Модель, которая позволяет определить вероятность появления слова в данном языке на основе анализа частоты и порядка появления морфем в данном языке.

Например, с помощью такой системы можно определить вероятность появления слова 皆既日食: P(皆既日食) = P(皆,既)*P(既,日)*P(日,食) (в случае использования диграммной модели).

Программная реализация таких моделей для языков с алфавитной письменностью затруднена из-за того, что одна и та же морфема может писаться разными символами: P(разукрашенный) = P(раз-)*P(раз-, у-) *P(у-, -крас-)*(-крас-, -енн-)*P(-енн-, -ый).

Следует отметить следующие преимущества: простота реализации для систем с иероглифической письменностью (китайский) и достаточно высокая точность работы.

3. Модель, охватывающая сразу несколько языковых уровней. Такая модель позволяет учитывать связи между языковыми единицами разных уровней. Графическим представлением данной модели является фрактальный граф.

ПРАКТИЧЕСКАЯ РЕАЛИЗАЦИЯ

Нами написана программа, которая позволяет построить модель определенного текста [4].

Программа оформлена как консольное приложение, так как такая реализация позволяет автоматизировать выполняемые программой действия

с помощью командных файлов. Результаты записываются в формате CVS, что позволяет использовать для анализа данных табличный редактор.

Программа вызывается помощью команды lmodgen. Ей передаются три параметра: имя файла описания алфавита языка; имя файла с текстом, на основании которого будет построена модель; и имя файла, в котором будет сохранён результат.

Все файлы должны быть в формате Unicode, в кодировке UCS2 с прямым порядком байтов.

Файл описания алфавита представляет собой обычный текстовый файл, каждая строчка которого представляет собой определённый символ в алфавите языка. Все варианты символа перечисляются в той же строке (подряд, без каких-либо разделителей). Файл текстом, по которому будет построена модель, также должен быть сохранён как обычный текст в формате Unicode (UCS2).

Результат будет сохранён как файл в формате CVS с двумя столбцами (сочетание трёх символов, частота) в кодировке UCS-2.

ЗАКЛЮЧЕНИЕ

В зависимости от поставленной задачи используются разные виды языковых моделей: для коррекции опечаток или распознавания голоса используются модели, работающие на уровне слов.

В своей работе мы написали программу, которая позволяет построить языковую модель на основании данных математической статистики и теории вероятностей. Программа позволит в будущем определить, насколько эффективен данный вид моделей, что требует дальнейшего исследования.

Литература

1. Интернет-адрес: http://language-model.narod.ru/full_report.doc.
2. *Зубов А. В., Зубова И. И.* Основы искусственного интеллекта для лингвистов. – М.: Университетская книга, Логос, 2007.
3. *Марчук Ю. Н.* Компьютерная лингвистика – М.: АСТ, Восток–запад, 2007.
4. Интернет-адрес: http://language-model.narod.ru/lmodgen.zip.

МЕХАНИЗМ ПРИНЯТИЯ РЕШЕНИЙ КЛЕТОЧНЫМ АВТОМАТОМ, ОСНОВАННЫМ НА НЕЧЁТКОЙ ЛОГИКЕ

Д. А. Кушнарёв, С. А. Филимонов

ВВЕДЕНИЕ

Наверное, самой впечатляющей способностью человеческого интеллекта является способность принимать правильные решения в условиях