Вывод и оценка параметров дальнодействующей триграммной модели языка

Протасов С. В.

МФТИ

Диалог 2008

Обзор доклада

1. Модель языка
   - Статистические модели языка
   - Для чего нужна модель языка?
   - Критерии оценки качества языковых моделей
   - Рейтинг языковых моделей
2. Дальнодействующие триграммные модели
   - Вероятностная модель грамматики связей
   - Исходные данные

Обзор доклада

Что такое модель языка?

Модель языка - это распределение вероятности различных последовательностей слов.

- Вероятность P("И ничего кроме правды") ≈ 0.0001
- Вероятность P("и ниче о рома да дым") ≈ 0

Всё, что Вам нужно знать о вероятности

P(X) значит вероятность, что X истина

- P(ребенок - мальчик) = 0.5 (процент мальчиков из общего количества)
- P(ребенка зовут Вася) = 0.01 (процент Вась из общего количества)

Модель языка

Модель дает вероятность последовательности слов

- P("И ничего кроме правды") ≈ 0.001
- P("и ничего о рома да дым") ≈ 0

Число последовательностей слов бесконечно, но полная сумма вероятностей равна единице.

Плохая модель языка

Плохая модель языка

- P("И ничего кроме правды") ≈ 0.001
- P("и ничего о рома да дым") ≈ 0.001

Плохая модель либо завышает вероятность некорректных предложений, либо занижает вероятность корректных.

Прикладные задачи

Возможные прикладные задачи

- Дикторонезависимое распознавание непрерывной речи.
- Статистический машинный перевод.
- Системы диалога на ЕЯ.
- Системы поиска информации.
- Оптическое распознавание (рукописного) текста.

Возможные прикладные задачи

Возможные прикладные задачи

- Кластеризация новостных сообщений по событиям.
- Коррекция опечаток (с использованием контекста).
- Нахождение соседних блоков в газетах.
- Сегментация текста на предложения.

Возможные прикладные задачи

Возможные прикладные задачи

- Идентификация автора.
- Идентификация спама в почте.
- Идентификация форума/блога.
- Идентификация языка текста.
- Борьба с поисковым спамом (дорвеи).
- Сжатие текста. Архиваторы.

Список далеко не полный....

Возможные прикладные задачи

Коррекция опечаток (с использованием контекста)

- Опечатка “езать” - это либо “ехать”, либо “резать”.
- езать хлеб - резать хлеб
- так как P(резать хлеб) $>$ P(ехать хлеб)
- езать в автобусе - ехать в автобусе
- так как P(ехать в автобусе) $>$ P(резать в автобусе)

Модель языка, используя контекст, исправляет опечатки более точно

Возможные прикладные задачи

Распознавание речи

Звук -> Кепстры -> Трифоны -> Текст

- На входе микрофона “И ничего кроме правды”
- Звуковой сигнал оцифровывается, преобразовывается и подается на вход акустической модели
- Акустическая модель выдает последовательность фонем, например “инечевооромэдафд”
- “инечевооромэдафд” может означать различные последовательности слов, например:
- “и ничего о рома да дым” ИЛИ “И ничего кроме правды”
- P(и ничего о рома да дым) < P(И ничего кроме правды)

Качественная модель языка позволяет значительно уменьшить число ошибок при распознавании слитной речи.

Возможные прикладные задачи

Борьба с дорвеями

Дорвеи - страницы поисковых спамеров, которые обманывают поисковики и стараются попасть в результаты поиска.

- Вероятность P(“Dvd cdrw3 usb Тип Б, 1 ru Все последние новости ноутбук acer 313 на laquoБитрикс Управление сайтомraquo Грамотность ноутбук acer 313 on line 160861227 Заказ on line out”) должна быть ≈ 0

Качественная модель языка может выявлять бессмысленные предложения и занижать рейтинг страниц “без смысла”.

Критерии оценки качества моделей

Критерии оценки качества моделей

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

Текущие лидеры по критерию кросс-энтропии

Текущие лидеры по критерию кросс-энтропии

- Человек - Игра Шеннона(1.3 бит на символ)
- Триграммная модель, основанная на классах (Brown 1992) 1.8 бит на символ.
- Кэш модели. Модели с памятью.(?)
- Контекстно-свободные модели. (?)

Обзор доклада

Вероятностная модель грамматики связей (Probabilistic Link Grammar Model)

Вероятностная грамматика связей (PLGM)

- Принадлежит к классу контекстно-свободных.
- Имеет эффективный алгоритм разбора (n^3 от числа слов)
- Возможны циклы, у связей нет направлений, связи могут иметь типы и иерархию.
- PLGM может быть натренирована, обучена на неразмеченном корпусе текстов.
- Дальнодействующая триграммная модель - частный случай PLGM.

Грамматика связей (Link Grammar)

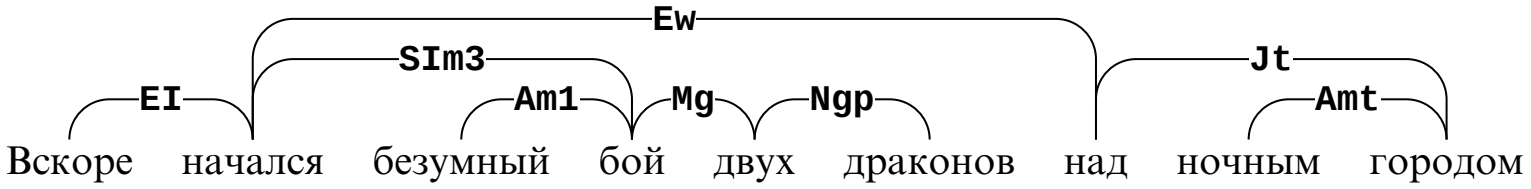


Рис. 1: *Связка слов.*

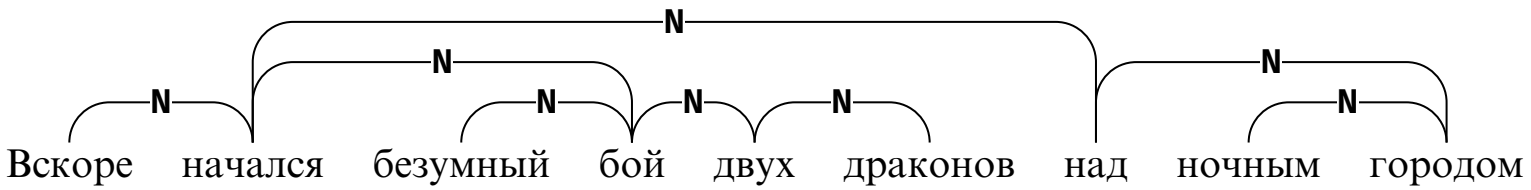
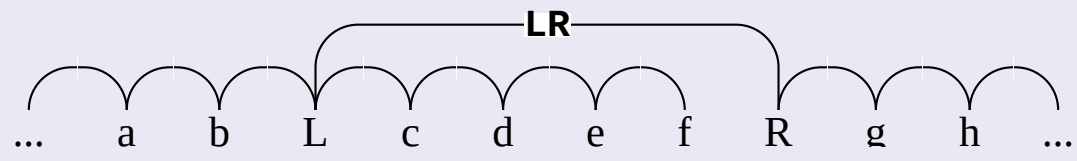


Рис. 2: *Единственный тип связи N.*

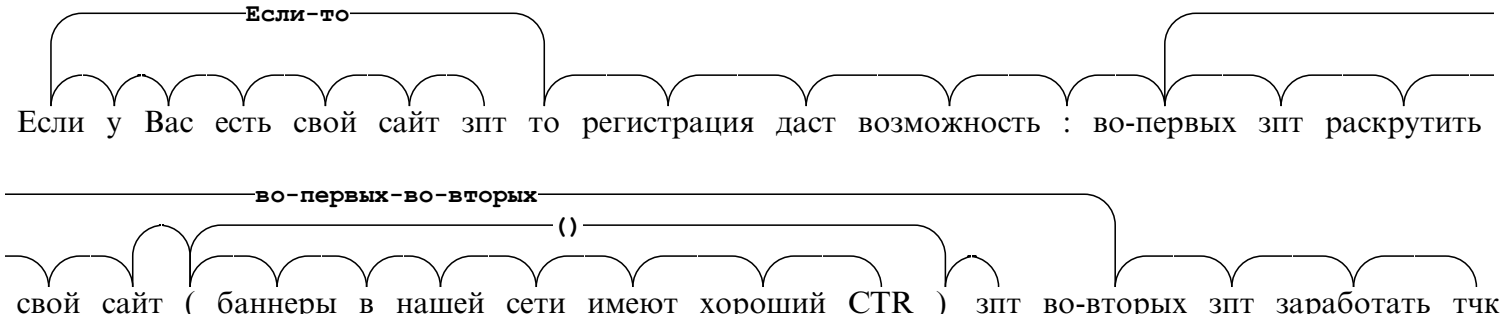
Дальнодействующие триграммы

Дальнодействующие триграммы



упрощенный вариант грамматики связей

Дальнодействующие триграммы



Исходные данные

Исходные данные для обучения

- Корпус из 11 млн неразмеченных предложений русского языка.
- 150 млн слов всего, 13 слов в среднем на предложение.
- 100 тыс различных слов. (Без учета морфологии)
- Словосочетания типа “вряд_ли” считаются за одно слово.
- Создан автоматически на базе интернет текстов.

Результаты

L	R	$log(Gain_{LR})$	$d_{(branch_{LR}\|L)}$	$d_{LR(halt)^{-1}}$
(	)	11.05	0.85	4.4
Если	то	8.81	0.35	7.1
либо	либо	8.44	0.33	4.2
"	"	8.33	0.21	7.9
Ни	ни	7.92	0.42	2.6
Чем	тем	7.78	0.44	5.0
столько	сколько	7.76	0.26	2.6
Чем_больше	тем	7.66	0.95	4.2
Что_касается	то	7.47	0.75	3.5
ни	ни	7.43	0.21	2.8

Таблица: Примеры пар слов (top10 словаря)

Результаты

Мусор ли это?

L	R	$log_{(Gain_{LR})}$	$d_{(branch_{LR}\|L)}$	$d_{LR(halt)^{-1}}$
аккорды	альбомы	7.24	0.63	10.8
мужчины	женщины	6.80	0.14	2.9
логин	пароль	6.56	0.69	1.4
товаров	услуг	6.55	0.19	1.8
финансы	маркетинг	6.21	0.28	8.1
плюсы	минусы	6.20	0.55	1.6
посолить	поперчить	5.86	0.37	1.1

Или данные пары образуют новые типы (синтаксических) связей?

Результаты

Результаты

- На первом шаге EM создан словарь парных слов.
- На первый взгляд словарь содержит мало мусора
- Что есть мусор? Нам неизвестно.

Планы

- Проделать несколько шагов EM (ожидания-максимизации).
- Запрограммировать сглаживание модели.
- Сравнить кросс-энтропию с триграммной моделью.
- Подключить в практическую задачу.

Спасибо! Вопросы?

Спасибо за внимание!

- Протасов Сергей Владимирович
- http://sz.ru/parser/
- Вопросы?