

Обучение с нуля грамматики связей для русского языка

Сергей Протасов

Факультет Управления и Прикладной Математики
МФТИ

КИИ-2006



Обзор доклада

- 1 Предыстория
 - Мотивация
 - Прикладные задачи
 - Анализаторы для русского языка
 - Виды грамматик
 - Критерии оценки качества
 - Рейтинг моделей
- 2 Алгоритм обучения грамматики
 - Родословная алгоритма
 - Грамматика связей
 - Исходные данные



Обзор доклада

- 1 Предыстория
 - Мотивация
 - Прикладные задачи
 - Анализаторы для русского языка
 - Виды грамматик
 - Критерии оценки качества
 - Рейтинг моделей
- 2 Алгоритм обучения грамматики
 - Родословная алгоритма
 - Грамматика связей
 - Исходные данные



Мотивация

Требуется программными методами получать ответы на следующие вопросы:

- **Верно ли** данное предложение с точки зрения русского языка?
- Какое предложение из списка **лучше всего звучит**?
- Где в данном предложении **самые важные слова**?
- **Какие связи** между словами в данном предложении?



Прикладные задачи

Возможные прикладные задачи

- Дикторонезависимое распознавание непрерывной речи.
- Подсистема проверки синтаксиса в текстовых редакторах.
- Системы диалога на ЕЯ.
- Системы поиска информации.
- Извлечение фактов из потока сообщений.
- Статистический машинный перевод.



Публичная демонстрация

<http://aot.ru>

opensource, выявление зависимостей

<http://rco.ru>

коммерческая, извлечение фактов

<http://sem1p.com>

коммерческая, определение роли слов в предложении.

<http://sz.ru/parser/>

opensource, выявление связей.

Все правила для грамматик созданы вручную!

Виды грамматик

Разновидности грамматик в лингвоанализаторах.

- Контекстно зависимые.
- Контекстно свободные.
- N-gram модели.



Удобство с точки зрения лингвистов

Удобство грамматик с точки зрения лингвистов

- Контекстно зависимые - очень удобно.
- Контекстно свободные - неудобно.
- N-gram модели - противоречит здравому смыслу (нет дальних связей).



Удобство с точки зрения программистов

Удобство грамматик с точки зрения программистов

- Контекстно зависимые - **комбинаторный взрыв**.
- Контекстно свободные - неудобно.
- N-gram модели - очень **легко**.



Временные затраты на анализ предложений

Временные затраты на анализ предложений

- Контекстно зависимые - экспонента Ce^n .
- Контекстно свободные - полином Cn^3 .
- N-gram модели - линейное время Cn .

n - число слов в предложении.



Критерии оценки качества анализаторов

Критерии оценки качества анализаторов

- Коэффициент правильных разборов.
- Коэффициент правильно установленных связей.
- Коэффициент правильно размеченных слов.
- Кросс-энтропия и перплексивность.



Текущие лидеры по критерию кросс-энтропии

Текущие лидеры по критерию кросс-энтропии

- Человек - Игра Шеннона (1.3 бит на символ)
- 3-gramm class-based модель (Brown 1992) 1.8 бит на символ.
- Контекстно-свободные модели.
- Контекстно-зависимые модели.



Обзор доклада

- 1 Предыстория
 - Мотивация
 - Прикладные задачи
 - Анализаторы для русского языка
 - Виды грамматик
 - Критерии оценки качества
 - Рейтинг моделей
- 2 Алгоритм обучения грамматики
 - Родословная алгоритма
 - Грамматика связей
 - Исходные данные



Алгоритм обучения

Родословная алгоритма

- EM (Expectation Maximization), один из методов максимального правдоподобия.
- IO (Inside-Outside) Algorithm, частный случай Expectation Maximization.
- Inside-Outside Link Grammar. Частный случай Inside-Outside.



Грамматика связей

Грамматика связей

- Принадлежит к классу контекстно-свободных (компромисс между программистами и лингвистами)
- Имеет эффективный алгоритм разбора (n^3 от числа слов)
- Не является грамматикой зависимостей. (возможны циклы, у связей нет направлений)



Грамматика связей

СВЯЗКИ СЛОВ

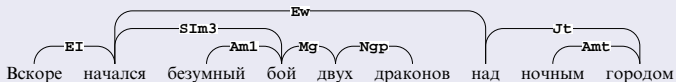


Рис. 1: Связка слов.

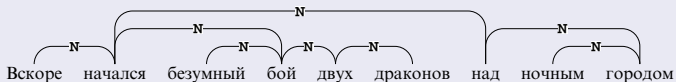


Рис. 2: Единственный тип связи N.

Исходные данные

Исходные данные для обучения

- Корпус из 2780 предложений русского языка.
- 355 разных высокочастотных слов.
- Длина предложения от 3 до 8 слов.



Результаты и Выводы

Результаты и Выводы

- Достигнуто улучшение кросс-энтропии по сравнению с биграмной моделью. 6 бит на слово
- Достигнут **хороший коэффициент разбора** предложений. 60 процентов приемлемых разборов.
- При увеличении корпуса обучения результаты **должны улучшаться..** Так как слова упрощаются.

Перспективы

- Увеличение скорости работы.
- Обучение на крупных текстах (100 тыс предложений).



Спасибо

Спасибо за внимание!

- Сергей Протасов
- parser@svp.zuzino.net.ru
- <http://sz.ru/parser/>