Компьютерный анализ естественно-языкового текста

Кафедра информационных систем в искусстве и гуманитарных науках

Компьютерный анализ естественно-языкового текста

СТРУКТУРА КУРСА

1. Введение в дисциплину
2. Автоматический анализ текста на морфологическом уровне
3. Автоматический анализ текста на синтаксическом уровне
4. Семантический компонент в системах автоматического анализа текста

СТРУКТУРА КУРСА

3. Автоматический анализ текста на синтаксическом уровне

   - *Задачи анализа текста на синтаксическом уровне*
   - *Модели представления структуры высказывания*
   - *Примеры реализации синтаксического анализа*

СТРУКТУРА КУРСА

3. Автоматический анализ текста на синтаксическом уровне

   – *Задачи анализа текста на синтаксическом уровне*
   – *Модели представления структуры высказывания*
   – *Примеры реализации синтаксического анализа*

ФОРМАЛЬНЫЙ ПОДХОД К ОРГАНИЗАЦИИ СИНТАКСИЧЕСКОГО АНАЛИЗА

- Мы хотим наши знания о синтаксисе формализовать. Каким метаязыком можно пользоваться?

ФОРМАЛЬНЫЙ ПОДХОД К ОРГАНИЗАЦИИ СИНТАКСИЧЕСКОГО АНАЛИЗА

- Мы хотим наши знания о синтаксисе формализовать. Каким метаязыком можно пользоваться?

- структуры составляющих
- структуры зависимостей
- гибридные модели

ФОРМАЛЬНЫЙ ПОДХОД К ОРГАНИЗАЦИИ СИНТАКСИЧЕСКОГО АНАЛИЗА

- Мы хотим наши знания о синтаксисе формализовать. Определились с метаязыком.
- А насколько этот метаязык способен отобразить наши знания о синтаксисе?

ФОРМАЛЬНЫЙ ПОДХОД К ОРГАНИЗАЦИИ СИНТАКСИЧЕСКОГО АНАЛИЗА

- Мы хотим наши знания о синтаксисе формализовать. Определились с метаязыком.
- А насколько этот метаязык способен отобразить наши знания о синтаксисе?

  Существуют описания (фрагментов) естественных языков, строящиеся на основе:

  -структур составляющих (ранние версии порождающей грамматики, …)

  -структур зависимостей (теория «Смысл⇔Текст, …)

  -гибридные структуры (поздние версии порождающей грамматики, …)

ФОРМАЛЬНЫЙ ПОДХОД К ОРГАНИЗАЦИИ СИНТАКСИЧЕСКОГО АНАЛИЗА

- Мы хотим наши знания о синтаксисе формализовать.
  Определились с метаязыком.
  Можем опереться на существующие описания (фрагментов) естественных языков – «грамматики»
- А как пользоваться этими описаниями для автоматической реализации синтаксического анализа?

ФОРМАЛЬНЫЙ ПОДХОД К ОРГАНИЗАЦИИ СИНТАКСИЧЕСКОГО АНАЛИЗА

- Мы хотим наши знания о синтаксисе формализовать.
  Определились с метаязыком.
  Можем опереться на существующие описания (фрагментов) естественных языков – «грамматики»
- А как пользоваться этими описаниями для автоматической реализации синтаксического анализа?

  Стоит вопрос о переходе от описания «что бывает в языке» к описанию алгоритма «как отождествить то, что видим в данном предложении, с тем, что бывает в языке»

ФОРМАЛЬНЫЙ ПОДХОД К ОРГАНИЗАЦИИ СИНТАКСИЧЕСКОГО АНАЛИЗА

- Мы хотим наши знания о синтаксисе формализовать. Определились с метаязыком.
  Можем опереться на существующие описания (фрагментов) естественных языков – «грамматики»
- А как пользоваться этими описаниями для автоматической реализации синтаксического анализа?

  Стоит вопрос о **парсинге**

  Процедура, которая предложению на некотором языке приписывает описание его структуры на специально предназначенном для этого метаязыке.

  Синоним в информатике – «синтаксический анализ» (также: «синтаксический разбор»)

ФОРМАЛЬНЫЙ ПОДХОД К ОРГАНИЗАЦИИ СИНТАКСИЧЕСКОГО АНАЛИЗА

- Мы хотим наши знания о синтаксисе формализовать.
  Определились с метаязыком.
  Можем опереться на существующие описания (фрагментов) естественных языков – «грамматики»

  ПАРСИНГ

  1) Для грамматик составляющих – проще (для некоторых классов – совсем просто)

  2) Для грамматик зависимостей – сложнее

  3) На практике – чаще гибридные структуры, используются алгоритмы с несколькими проходами по предложению, большое количество решений для частных случаев (АОТ)

ПАРСИНГ: ГРАММАТИКИ СОСТАВЛЯЮЩИХ

неограниченные грамматики	
контекстно-зависимые грамматики (грамматики НС)	
контекстно-свободные грамматики	Соответствуют обычному пониманию структур составляющих
автоматные (регулярные) грамматики	Соответствуют частному случаю в обычном понимании структур составляющих

ПАРСИНГ: ГРАММАТИКИ СОСТАВЛЯЮЩИХ

неограниченные грамматики	Их структурный коррелят значительно сложнее структур составляющих в обычном понимании
контекстно-зависимые грамматики (грамматики НС)	Их структурный коррелят сложнее структур составляющих в обычном понимании
контекстно-свободные грамматики	Соответствуют обычному пониманию структур составляющих
автоматные (регулярные) грамматики	Соответствуют частному случаю в обычном понимании структур составляющих

ПАРСИНГ: ГРАММАТИКИ СОСТАВЛЯЮЩИХ и АВТОМАТЫ

неограниченные грамматики	машины Тьюринга
контекстно-зависимые грамматики (грамматики НС)	линейно ограниченные автоматы / машины Тьюринга с конечной лентой
контекстно-свободные грамматики	автоматы с магазинной памятью (стековые автоматы)
автоматные (регулярные) грамматики	конечные автоматы

ПАРСИНГ: ГРАММАТИКИ СОСТАВЛЯЮЩИХ и СЕТИ ПЕРЕХОДОВ

неограниченные грамматики	усиленные (=расширенные) сети переходов
контекстно-зависимые грамматики (грамматики НС)	
контекстно-свободные грамматики	рекурсивные сети переходов
автоматные (регулярные) грамматики	базовые сети переходов (=диаграммы переходов конечных автоматов)

ПАРСИНГ: ГРАММАТИКИ СОСТАВЛЯЮЩИХ и АВТОМАТЫ

…	…
контекстно-свободные грамматики	автоматы с магазинной памятью (стековые автоматы)
автоматные (регулярные) грамматики	конечные автоматы

ПАРСИНГ для КОНТЕКСТНО-СВОБОДНЫХ ГРАММАТИК

Для КС грамматик нет универсального алгоритма/процедуры перехода «Грамматика → Автомат»

Тем не менее, автомат – это не единственная форма задания алгоритма парсинга; для более общей задачи создать алгоритм перехода «Грамматика → Парсинг» существуют универсальные решения и в классе КС грамматик

Однако эти универсальные решения, т.е. способы по любой КС грамматике построить алгоритм парсинга, малоэффективны, т.к.

- состоят из нескольких этапов
- тот алгоритм парсинга, который получается в результате такой универсальной процедуры, слишком затратный в отношении вычислительных ресурсов

Для некоторых классов КС-грамматик (но не для всех) существуют более эффективные способы организовать парсинг

ПАРСИНГ для КОНТЕКСТНО-СВОБОДНЫХ ГРАММАТИК

Наиболее известные универсальные способы построения по любой КС грамматике алгоритма парсинга:

- алгоритм Кока-Янгера-Касами
- алгоритм Эрли

Оба предусматривают в качестве промежуточного шага построение вспомогательной структуры данных (таблица для алгоритма К-Я-К, список для алгоритма Эрли)

Оба включают в качестве входа не только грамматику, но и конкретное разбираемое предложение

Оба требуют времени разбора n^3 и объема затрачиваемой памяти n^2, где n – длина разбираемого предложения (хотя для некоторых подтипов КС-грамматик алгоритм Эрли может работать затрачивать линейные время и объем памяти) .

ПАРСИНГ для КОНТЕКСТНО-СВОБОДНЫХ ГРАММАТИК

Алгоритм Кока-Янгера-Касами (пример)

Дано:

1. грамматика

   S → NP VP (1)

   NP → Det N (2)

   VP → V NP (3)

   N → boy | ball (4) (5)

   Det → the (6)

   V → sees (7)

2. предложение ***the boy sees the ball***

ПАРСИНГ для КОНТЕКСТНО-СВОБОДНЫХ ГРАММАТИК

Алгоритм Кока-Янгера-Касами (пример)

Этапы:

1. Построение таблицы

	t_{15}				
	t_{14}	t_{24}			
	t_{13}	t_{23}	t_{33}		
	t_{12}	t_{22}	t_{32}	t_{42}	
	t_{11}	t_{21}	t_{31}	t_{41}	t_{51}

2. Разбор по таблице

ПАРСИНГ для КОНТЕКСТНО-СВОБОДНЫХ ГРАММАТИК

Алгоритм Кока-Янгера-Касами (пример)

Принцип построения таблицы:

	t_{15}				
	t_{14}	t_{24}			
	t_{13}	t_{23}	t_{33}		
	t_{12}	t_{22}	t_{32}	t_{42}	
	t_{11}	t_{21}	t_{31}	t_{41}	t_{51}

В клетки t_{ij} вносятся такие нетерминальные символы ***A*** (левые части правил грамматики), что из A можно вывести j слов разбираемого предложения, начиная с i-го слова.

ПАРСИНГ для КОНТЕКСТНО-СВОБОДНЫХ ГРАММАТИК

Алгоритм Кока-Янгера-Касами (пример)

Построение таблицы для данного примера:

	S				
	——	——			
	——	——	VP		
	NP	——	——	NP	
	Det	N	V	Det	N

Далее – разбор по таблице…

ПАРСИНГ: ГРАММАТИКИ ЗАВИСИМОСТЕЙ

…

- (*не входит в данный курс*)

…

Более подробная информация об организации парсинга для структур зависимостей

(на английском языке)

http://bulba.sdsu.edu/cl/Members/rmalouf/courses/ling-795-dependency-parsing

http://aclweb.org/mirror/acl2006/program/tutorials/dependency.html

ПАРСИНГ: ПРИМЕР ДЛЯ ГИБРИДНОЙ МОДЕЛИ СИНТАКСИСА (АОТ)

- Синтаксический процессор ДИАЛИНГ (Л. Гершензон, Д.Панкратов, А.Сокирко) разработан в 1998-2001 г. на основе процессора ПОЛИТЕКСТ (система анализа политических текстов Центра информационных исследований).
- Используется понятие синтаксических групп
- На входе результаты работы графематического и морфологического модуля (каждая словоформа представлена множеством морфологических омонимов)

ПАРСИНГ: ПРИМЕР ДЛЯ ГИБРИДНОЙ МОДЕЛИ СИНТАКСИСА (АОТ) - 2

- Особенность архитектуры: двунаправленное взаимодействие модуля **сегментации** (=фрагментации, разбиение на предикативные единицы типа простых предложений) и **синтаксиса** (построения синтаксических групп слов в предложении).
- Перед анализом не ставится цель построить полную синтаксическую структуру (только объединяет в группы то, что можно объединить).
- Демонстрация анализа в режиме он-лайн: http://www.aot.ru/demo/synt.html (а также модуль SynAn пакета Dialing, загружаемого с сайта АОТ)

ПАРСИНГ: ПРИМЕР ДЛЯ ГИБРИДНОЙ МОДЕЛИ СИНТАКСИСА (АОТ) - 3

Этапы работы синтаксического анализатора

1. Первичная сегментация по пунктуации и сочинительным союзам с учетом простейших рядов однородных членов
2. Объединение элементов аналитических форм глагола
3. Выделение терминологических именных групп
4. Обработка существующих и восстановление пропущенных тире в функции связки
5. Построение множества МИ внутри сегментов
6. Объединение сочиненных сегментов

...

ПАРСИНГ: ПРИМЕР ДЛЯ ГИБРИДНОЙ МОДЕЛИ СИНТАКСИСА (АОТ) - 4

Этапы работы синтаксического анализатора

…

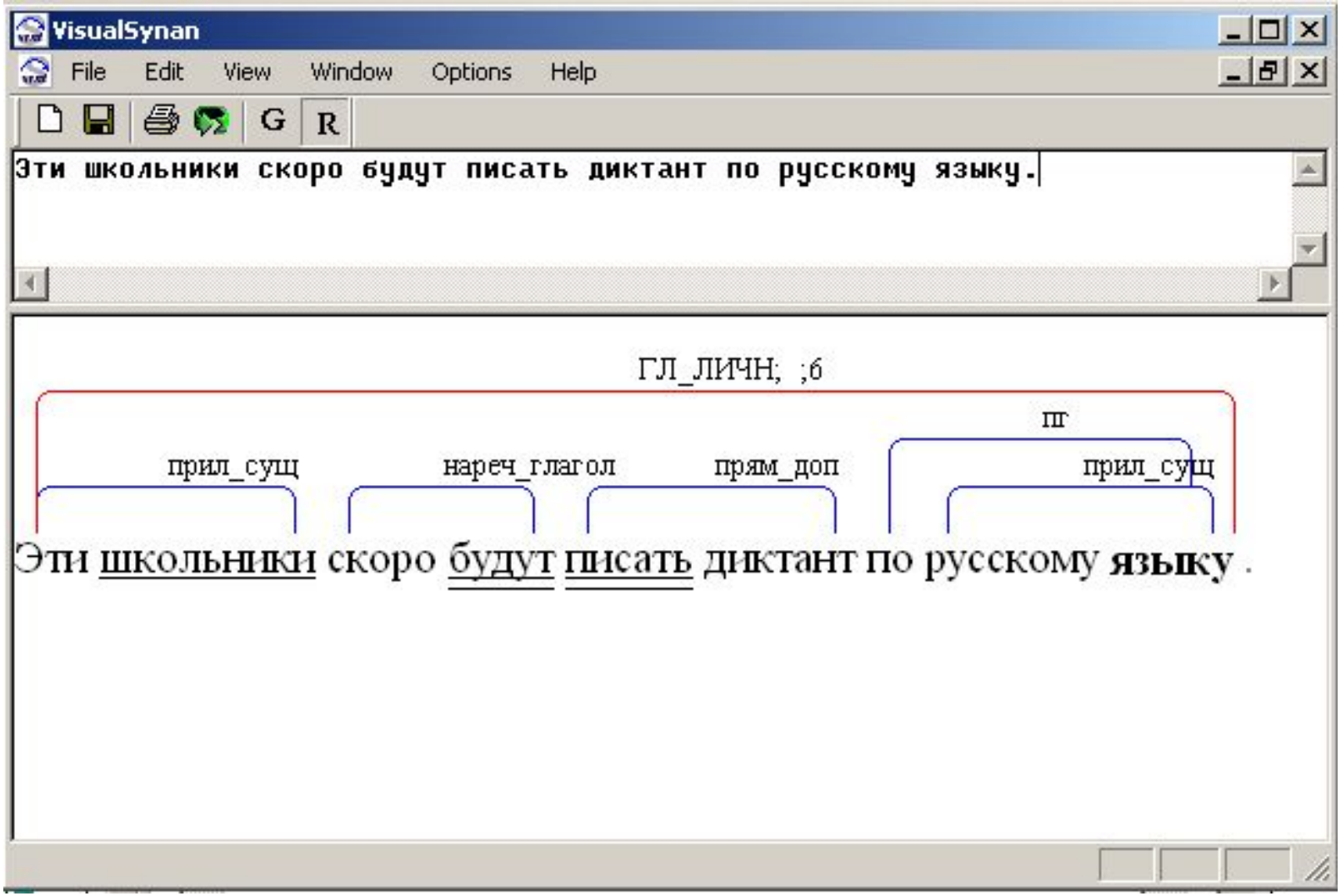
7. Построение сочиненных групп (именных, глагольных) внутри сегментов
8. Вложение сегментов (установление отношений подчинения)
9. Построение синт. групп, включающих вложенные сегменты
10. Объединение разрывных сегментов
11. Построение групп с использованием всех правил обработки МИ
12. Ранжирование МИ по синтаксическому покрытию

ПАРСИНГ: ПРИМЕР ДЛЯ ГИБРИДНОЙ МОДЕЛИ СИНТАКСИСА (АОТ) - 5

39 типов синтаксических групп, в том числе:

Тип	Название	Пример
Количественная группа (последовательность числительных)	КОЛИЧ	двадцать восемь
Последовательность чисел	СЛОЖ-ЧИСЛ	12,3, II-III
Группа существительного, пре-модифицированная одним или несколькими прилагательными	ПРИЛ-СУЩ	длинная тяжелая дорога, двигающийся человек
Группа существительного, пре-модифицированная наречным числительным	НАР-ЧИСЛ-СУЩ	много ребят, мало стульев
Группа существительного, пре-модифицированная числительным	СУЩ-ЧИСЛ	восемь попугаев, два человека
Предложная группа	ПГ	в дом, на холме
…		

ПАРСИНГ: ПРИМЕР ДЛЯ ГИБРИДНОЙ МОДЕЛИ СИНТАКСИСА (АОТ) - 6



РЕКОМЕНДОВАННАЯ ЛИТЕРАТУРА

- *Тестелец Я. Г.* Введение в общий синтаксис. М., 2001. (Главы I, II)
- *АОТ:* Синтаксический анализ. http://www.aot.ru/docs/synan.html
- *Ножов И.М.* Морфологическая и синтаксическая обработка текста (модели и программы). Дисс. … канд. тех. наук. М., 2003. http://www.aot.ru/docs/Nozhov/chapter3.pdf (Глава 3, I) или http://www.aot.ru/docs/Nozhov/msot.pdf. (диссертация полностью)

ДОПОЛНИТЕЛЬНАЯ ЛИТЕРАТУРА

- *Мельчук И. А.* Опыт теории лингвистических моделей «Смысл⇔Текст». М., 1974 (1999) (Глава II, § 1, 2)
- *Ахо А.*, *Ульман Дж.* Теория синтаксического анализа, перевода и компиляции. Том 1. Синтаксический анализ. М.: Мир, 1978.