ОПЫТ ИСПОЛЬЗОВАНИЯ СЕМАНТИКО-СИНТАКСИЧЕСКОГО АНАЛИЗАТОРА SEMSIN ДЛЯ СИНТЕЗА РУССКОЙ РЕЧИ

Чемерилов В. В., Савинов А.П.
Томский политехнический университет
vchemerilov@gmail.com

Введение

В 2009 году Б.М. Лобановым был проведен эксперимент с синтаксическим анализатором ЭТАП-3 с целью его адаптации для систем синтеза речи [1]. Несмотря на успешное прохождение всех тестов, ЭТАП-3 не смог полностьюразрешить вопросы омонимии, так как не содержал семантической информации при анализе текста. Семантико-синтаксический анализатор Semsin в отличии от ЭТАП-3 проводит не только синтаксический и морфологический анализы текста, но и семантический. В данной работе был проведен эксперимент с анализатором Semsin с целью его применения в русском речевом синтезе.

Анализатор Semsin

Semsin – это семантико-синтаксический анализатор, в задачи которого входит снятие частеречной и морфологической омонимии, построение синтаксического дерева зависимостей и частичное снятие лексической неоднозначности [2].

Принцип работы синтаксического анализатора Semsin:

1. Выделение токенов. Каждый абзац поданного на вход текста подвергается предварительному анализу с выделением минимальных, линейных и неделимых компонент текста – токенов.
2. Обработка отдельных слов. Разрешение морфологической омонимии засчет анализа ближайшего окружения.
3. Выделение групп с фамилиями, названиями, числами.
4. Подключение прилагательных и причастий, снятие неоднозначностей прилагательное–существительное. Поиск согласованного существительного.
5. Поиск предлогов на основе анализа предложных групп.
6. Сегментация текста по знакам препинания с целью нахождения предикативной вершины (обычно это глагол).
7. Поиск составных сказуемых на основе словарей актанктных глаголов.
8. Построение дерева синтаксических связей слов в предложении.

Особенности сематико-синтаксического анализатора Semsin

Semsin содержит 177 тысяч лексем, распределенным по 1660 классам, что позволит преобразовать даже сложный технический текст в качественную речь. В 2014 году были проведены испытания парсера Semsin с целью категоризация текстов для структурирования массива исторических документов [3]. В результате точность определения лемм была не ниже 97%. Semsin содержит более 20 тысяч названий и собственных имен, что позволит решить вопрос омонимии при синтезе речи.

В настоящее время база содержит более 4100 фразеологизмов и играет важную роль в снятии неоднозначности, особенно для составных предлогов, союзов и наречий [4]. Используется также отдельная база предлогов, хранящая классы существительных, с которыми они взаимодействуют, и названия связей с хозяевами предложных групп.

Исследование парсера SemSin

Для исследования возможностейсинтаксического анализатора Semsin на практике был проведен эксперимент. Была сделана выборка частей текста из пяти разных мест трех книг (предметная область – экономика). Части текста были записаны в отдельные текстовые файлы (по пять файлов на каждую книгу) и каждый файл был обработан Semsin. В результате работы была получена следующая статистика.

Таблица 1. Неоднозначность лемм Semsin(%).

	1	2	3	4	5	сред
Баскакова, Сейко — Экономика предприятия	1	0	0,9	0	0	0,44
Фролова — Экономика предприятия. Конспект лекций	0	0	0	0	0	0
Землянская, Эглит — Тезаурус по экономике предприятия	0	0	0	0	2,20	0,44

Таблица 2. Неоднозначность морфологии Semsin (%).

	1	2	3	4	5	сред
Баскакова, Сейко — Экономика предприятия	0	2,17	9, 01	11,67	4,23	5,42
Фролова — Экономика предприятия. Конспект лекций	8,51	7,46	24, 68	0	4,55	9,04
Землянская, Эглит — Тезаурус по экономике предприятия	13,33	12, 82	18,37	4, 17	9,89	11,72

Таблица 3. Время работы Semsin (слов в секунду).

	1	2	3	4	5	сред
Баскакова, Сейко — Экономика предприятия	70,4	50,82	60,83	54,86	46,01	56,6
Фролова — Экономика предприятия. Конспект лекций	47,84	57,26	55,5	60,85	48,69	54
Землянская, Эглит — Тезаурус по экономике предприятия	49,08	53,14	47,65	59,86	63,4	54,6

Парсер Semsin подобрал лемму практически к каждому слову в представленных частях текста. В 8,73 % случаях парсер не смог определить морфологические характеристики слова. Средняя скорость работы парсера 55 слов в секунду.

Синтаксический анализ Semsin:

1. Определяет связь между соседними словами в тексте с помощью вопроса, причем рассматривает каждое слово как часть речи (морфология), а не как член предложения (подлежащее, сказуемое, определение, дополнение, обстоятельство).

2. Выделяет подлежащее и сказуемое – связь субъект и объект между ними.

3. Выделяет деепричастия и причастия.

4. Если связи с соседними словами нет (заканчивается одна часть сложносочиненного предложения, начинается другая), то ставит знак “сочинит” – сочинительная связь (не смотря на то, что связь может быть подчинительной).

Парсер Semsin не определяет:

- Вид предложения (простое, сложное).
- Вид простого предложения (полное, неполное, распространенное, нераспространенное).
- Вид сложного предложения (сложносочиненное, сложноподчиненное и т.д.).
- Грамматическую основу предложения (подлежащее, сказуемое) и второстепенные члены предложения (определение, дополнение, обстоятельство).

Заключение

При интегрировании парсера Semsin в синтезатор речи, у синтезатора речи появятся следующие возможности:

- Снятие омографии словоформ. Зная лемму и семантический класс слов предложения можно определить на какую букву будет ставиться ударение, если предложение содержит омограф (слово, которое читается по-разному, но пишется одинаково).
- Синтагматический анализ предложения (деление предложения на синтагмы), который позволит определить:
  - Места пауз в предложении.
  - Места расстановки логического, фразового и синтагматических ударений.
- Актуальное членение предложения – выделие темы и ремы в предложении. Рема обычно выделяется фразовым ударением. Тема и рема разделяются паузой.
- Выделение интонационных конструкций (ИК).
- Определение и задание интонации перечисления, пояснения и сопоставления.
- Задание скорости произношения слова.
- Расшифровка сокращений.

Список использованных источников

1. Многоцелевой лингвистический процессор ЭТАП-3 [Электронный ресурс]. - URL: http://iitp.ru/ru/science/works/452.htm (дата обращения: 04.10.2016).

2. Каневский Е.А., Боярский К.К Предсинтаксический модуль в анализаторе SemSin [Электронный ресурс]. - URL: http://ojs.ifmo.ru/index.php/IMS/article/viewFile/46/47 (дата обращения: 04.10.2016).

3. Артемова Г.В., Боярский К.К., Гусарова Н.Ф., Добренко Н.В. Категоризация текстов для структурирования массива исторических документов [Электронный ресурс]. - URL:

4. http://rcdl.ru/doc/2014/paper/RCDL2014_159-164.pdf (дата обращения: 04.10.2016).

5. Каневский Е.А., Боярский К.К. Семантико-синтаксический анализатор SemSin [Электронный ресурс]. - URL: http://www.dialog-21.ru/digests/dialog2012/materials/pdf/Kanevsky.pdf (дата обращения: 04.10.2016).