ОШИБКИ И НЕДОСТАТКИ МОРФОЛОГИЧЕСКОГО И СЕМАНТИЧЕСКОГО АНАЛИЗОВ СЕМАНТИКО-СИНТАКСИЧЕСКОГО АНАЛИЗАТОРА ТЕКСТА SEMSIN

В.В. Чемерилов, А.С. Фадеев
Томский политехнический университет
vchemerilov@gmail.com

Введение

Синтезаторы речи по тексту – это системы, которые конвертируют текст, введенный пользователем, в искусственную речь. Как правило, такие системы проводят предварительную обработку текста с целью расставления маркеров интонации (пауз, ударений и т.д.). Так как большинство правил выделения интонации используют данные морфологического, синтаксического и семантического анализов текста, то в блоке обработки текстовых данных стали использовать специальные анализаторы [1]. В данной работе описаны ошибки и недостатки морфологического и семантического анализов парсера Semsin, которые выявились на этапе тестирования анализатора для его использования в системе синтеза русской речи.

Анализатор Semsin

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

Принцип работы анализатора Semsin состоит в следующем: на вход анализатора подается текст на русском языке, текст делится на абзацы. Каждый абзац подвергается предварительному анализу с выделением отдельных токенов [3]. После токенизации текст разбивается на предложения. Проводится обработка отдельных слов, выделяются группы слов с фамилиями, названиями, числами. На четвертом этапе работы системы подключаются прилагательные и причастия, происходит снятие неоднозначностей прилагательное–существительное. Подключаются предлоги, образуются предложные группы. После проведения сегментации (работа с обособленными оборотами, придаточными предложениями и т. д.) и поиска составных сказуемых и подлежащих, система строит синтаксическое дерево зависимостей и указывает типы синтаксических связей элементов дерева.

Семантический анализ Semsin

В качестве исходных данных, на основе которых проводится семантический и морфологический анализы парсера Semsin, используется словарь и классификатор В.А.Тузова [4]. Словарь Тузова основан на семантическом словаре С.А. Кузнецова [5] и морфологическом словаре А.А. Зализняка [6]. Исходный семантический словарь состоит из 177 тысяч лексем, распределенным по 1660 семантическим классам [2]. Около 14% слов содержат несколько семантических классов (имеют две или более лексем). Также словарь содержит более 40 тысяч названий (городов, сел, фирм, рек, учреждений и т.д.) и собственных имен.

Морфологический анализ Semsin

Для проведения морфологического анализа текста на основе морфологического словаря была разработана морфологическая база данных. В базе к каждой лексеме был подобран набор морфологических характеристик, номер своего класса и актанты вызываемых ею лексем. Данные были представлены в виде падежей или предлогов с соответствующими падежами, например: вПред, вВин и т.д. Также перед актантом указывались допустимые классы слов, которые могли их замещать. В том случае, если несколько лексем имели одно и о же описание, они объединялись в одну лексему.

Морфологический словарь содержит слова, которые могут имеют две или более
лексемы [3] – эти слова имеют различные морфологические характеристики, например: «печь» – это существительное женского рода, единственного числа или несовершенный глагол. Также морфологические характеристики лексем могут совпадать, но семантические классы различаться, например, слово «коса» имеет три семантических класса:

1. Жизнь, части живого, голова, волосы;
2. Физический объект, неодушевленный, вещь, утварь, инвентарь, с/х;
3. Физический объект, природа, природные зоны, ландшафт, берег.

Однако морфологические характеристики лексем совпадают: существительное, женский род, единственное число и т.д.

В семантико-синтаксическом анализаторе Semsin морфологический словарь реализован в виде таблицы Excel. Она содержит более 170 тысяч строк [2]. Некоторые из этих строк соответствуют не одной, а нескольким лексемам с одинаковой морфологией. Данная таблица используется морфологическим анализатором для проведения морфологического анализа слова.

Для решения проблемы устойчивых словосочетаний, анализатор использует специальную таблицу фразеологизмов, которая проводит разбор трех различных типов словосочетаний: полностью изменяемых (море по колено), частично изменяемых (дым коромыслом) и неизменяемых (в ту пору, а именно). На данный момент таблица содержит более 4800 фразеологизмов [3]. Она играет важную роль в снятии неоднозначности.

Также важным элементом морфологического анализа парсера Semsin является таблица предлогов, состоящая из 2200 сочетаний классов существительных, с которыми взаимодействуют предлоги.

Морфологический анализатор [2] проводит анализ входных данных – каждого слова в тексте. Результатом анализа является лемма слова с его морфологическими характеристиками (часть речи, род, число, падеж и т.д.).

Ошибки и недостатки морфологического и семантического анализов Semsin

В процессе тестирования семантико-синтаксического анализатора Semsin для его дальнейшего использования в системах синтеза русской речи, возникли ошибки, связанные с морфологическим и семантическим анализами парсера. Также в некоторых случаях анализатор представил недостаточно данных для выделения мест расстановки интонации. Для полноценного морфологического анализа (чтобы на его основе выделять места расстановки интонации) не хватает определения:

1. Качественных прилагательных, наречий и существительных;
2. Отглагольных существительных;
3. Отыменных прилагательных;

Semsin не определяет разряд наречия (только часть речи). Из ошибок морфологического анализа Semsin можно выделить неправильное определение морфологических характеристик. Данная ошибка возникает из-за неправильного подбора лексемы, например:

Ты сегодня ела кашу? Нет, слесарь.

Для ключевого слова ответа «слесарь», Semsin подобрал лемму «слесарить» и морфологические характеристики: глагол, несовершенный, непереходный и т.д. Хотя в данном контексте «слесарь» - существительное. В результате интонационный центр упал на слово вопроса «ела», хотя должен был упасть на слово «ты».

Одной из важнейших ошибок, связанных с семантическим анализом парсера Semsin, которая влияет на интонационный анализ текста, является то, что он не всегда правильно разрешает омонимию. Задача разрешения графической омонимии – одна из важных задач в синтезе русской речи [7]. Установив точное значение
слова (в данном случае семантический класс слова) появляется возможность определить его ударный слог, например: в слове «замок» в значении «Физический объект, неодушевленный, вещь, утварь, инструменты замки» ударение падает на второй слог, а в значении «Физический объект Поселения Постройка Жилье Дом» на первый. В некоторых случаях, Semsin неправильно определяет семантический класс омографа, а значит неправильно определяет ударный слог. Например, в некоторых случаях Semsin определяет имя «Маша» (в начале предложения) как деепричастие «машу» (Действие Труд Физический Махание). Также возникает ошибка, когда Semsin повторяет семантические классы одного и того же слова. От этой ошибки легко избавится, проведя фильтрацию семантических данных, но все же она имеет место быть.

Так как Semsin пытается самостоятельно разрешить омонимию, то в некоторых случаях, он выдает неполную информацию о семантическом анализе слова, например:

Олег пил сегодня чай? Да, чай.

Вопросительное слово ответа «чай» содержит следующие данные по семантическому анализу: «Вероятность». Однако, данное слово имеет еще один семантический класс «Жизнь Пища Напитки Безалкогольные». Если бы данные семантического анализа ключевого слова ответа были
полные (содержали бы оба семантических класса), то сопоставляя эти классы с семантическими классами слов вопроса, алгоритм автоматического выделения ударного слова в вопросительном предложении [8] выделил интонационный центр в слове «чай» (в вопросительном предложении). Однако в силу того, что Semsin выдал неполные данные по ключевому слову ответа, алгоритм не смог решить поставленной ему задачи.

Заключение

При соответствующей доработке семантико-синтаксического анализатора Semsin, выходные данные парсера можно будет использовать для разработки блока автоматического выделения интонации в лингвистическом процессоре при синтезе русской речи.

Список использованных источников

1. Иомдин Л.Л. Говорящий «ЭТАП». Опыт использования синтаксического анализатора системы ЭТАП в русском речевом синтезе /Л.Л. Иомдин, Б.М. Лобанов, Ю.С. Гецевич // Труды Международной конференции «Компьютерная лингвистика и интеллектуальные технологии», 2011. – С. 269-279.
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: 29.05.2017).
3. Каневский Е.А., Боярский К.К. Семантико-синтаксический анализатор Semsin // Научно-технический вестник информационных технологий механики и оптики. – 2015. – Т. 15 № 5. – С. 869-876.
4. Тузов В.А. Компьютерная семантика русского языка. – СПб.: Изд-во СПбГУ, 2004. –400 с.
5. Кузнецов С.А. Большой толковый словарь русского языка. – СПб.: Изд-во Норинт, 1998. – 1536 с.
6. Зализняк А.А. Грамматический словарь русского языка. Словоизменение. – М.: Русский язык, 1980. — 880 с.
7. Лобанов Б. М., Житко В. А. О решении задач снятия омонимии при распознавании и синтезе речи [Электронный ресурс]. - URL:https://libeldoc.bsuir.by/handle/123456789/4372 (дата обращения: 01.06.2017).
8. Чемерилов В.В., Фадеев А.С., Мишунин О.Б. Алгоритм автоматического выделения интонационного центра в вопросительном предложении без явного вопросительного слова на основе семантических связей предложений при автоматическом анализе текста для синтеза русской речи // Современные наукоемкие технологии. – 2017. – № 11. – С. 75-79.