Опыт построения полной морфемно-ориентированной семантической сети для русского языка

А. И. Пацкин
РосНИИ Искусственного интеллекта, Москва
aleksandr@tochka.ru

Учет словообразовательных связей и морфемной структуры русских слов может дать значительный выигрыш при построении семантических сетей в объеме всего языка. Объем необходимых семантических описаний и, соответственно, работы компьютерных лексикографов при приоритетном описании морфем перед лексемами может сократиться на порядок, учитывая среднюю производительность русского корня равную примерно 13 (по А.Н. Тихонову). А с учетом того факта, что наиболее частотные корни одновременно и наиболее производительные (мощность "нести" равна 540), трудно переоценить эффективность приоритетного семантического описания морфем, для получения компактного описания семантики, используемой в русском ЕЯ-анализе. В РОСНИИ ИИ с 2001-го года автором ведется работа по созданию гиперсловаря "Ариадна", содержащего морфемную, словообразовательную и словоизменительную информацию по русскому языку в объеме словника словаря А.А. Зализняка. Назначение данного гиперсловаря - служить основой (каркасом) для семантической сети, предназначенной для глубокого смыслового анализа русского текста. Гиперсловарь строится на базе системы представления знаний Абриаль 2. Текущее состояние данной работы отражено в докладе.

Сокращения в тексте

- **СЗ** - Грамматический словарь русского языка А.А. Зализняка [2]
- **СКЕ** - Словарь морфем русского языка А.И. Кузнецовой и Т.Ф Ефремовой [3]
- **СТ** - Словообразовательный словарь русского языка А.Н. Тихонова [7]

История проекта

Историю проекта "Ариадна" можно разделить на три этапа. Первый из них был выполнен автором осенью 2001-года. Целью его была проверка пригодности разработанной автором системы Абриаль, для поддержки больших баз знаний по русскому языку, основанных на материале стотысячного Словаря Зализняка (**СЗ**)[2]. Первоначально планировалось только реализовать зализняковскую грамматику (описанную в предисловии к словарю и ставшую сейчас "классикой"). Что и было сделано. Однако, в связи с чисто технической трудностью, (а именно, неэффективностью доступа к большим массивам словарных элементов через интерфейс имевшейся на тот момент версии программы Абриаль), было решено по возможности превратить сплошной указатель словаря в дерево, для чего все слова из словаря разделить на сегменты. Однако механически делить слова на части показалось неинтересным.

Гораздо более заманчивым заделом на будущее представлялось использовать в некотором смысле "правильное" разбиение слов на морфемы. Для проведения морфемной сегментации был сформирован список корней и аффиксов из Словаря морфем Кузнецовой и Ефремовой [3] (**СКЕ**). К нему были добавлены наиболее частотные корневые морфемы, не вошедшие в **СКЕ**. На основе этих списков и написанной автором эвристической процедуры была проведена морфемная сегментация слов из СЗ, которая дала достаточно хорошие результаты, т.е. правильно разбила слова в основном на массиве исконной русской лексики. Например, слова типа *за(вод)о-у(прав)лени-е* или *(сноп)о(вяз)алк-а* разбились почти все правильно, но правильно лишь в рамках этимологически ориентированного подхода СКЕ, где в слове *завод* корень *вод* а в слове *ответ* - корень *вет*. (А эта методология имеет альтернативы, в частности для семантически ориентированного подхода А.Н. Тихонова в слове *подбородок* корень *подбородок*). В результате этого первого этапа была создана словарная сетевая база знаний - **Гиперсловарь Ариадна**. Для этого Гиперсловаря средствами системы Абриаль была сделана гипертекстовая оболочка, т.е. интерфейс для навигации в веб-стиле, через который Гиперсловарь просматривался как виртуальный сайт. Работа была представлена на конференции ДИАЛОГ-2002 [5].

Затем в работе над Гиперсловарем произошел годовой перерыв, необходимый для доработки системы Абриаль, которая за это время превратилась в Абриаль 2 - полноценную систему разработки визуальных приложений для сетевых баз знаний. В итоге получилась значительно более наглядная и удобная интерфейсная оболочка для Гиперсловаря, позволяющая неподготовленному пользователю осуществлять навигацию по сложной сети словарных объектов, произвольно меняя направление, и изменять попутно обозреваемые данные, что могло быть полезным как для работы конструктора семантической сети, так и для лингвистических исследований.

Ниже показан фрагмент страницы интерфейса программы, в котором присутствуют ошибочные разбиения, отмеченные (только здесь, естественно) *курсивом*:

Суффиксный сегмент *'ативн'* входит в слова				
агглютин|ативн|ый	**Ассоци|ативн|ый**	**информ|ативн|ый**	**кауз|ативн|ый**	**коммуник|ативн|ый**
коммут|ативн|ый	**компар|ативн|ый**	**конспир|ативн|ый**	***конф|е|дер|ативн|ый***	***ко|о|пер|ативн|ый***
нег|ативн|ый	**норм|ативн|ый**	***о|пер|ативн|ый***	***порт|ативн|ый***	**рекупер|ативн|ый**
с|верх|норм|ативн|ый	**сепар|ативн|ый**	**факульт|ативн|ый**	**федер|ативн|ый**	

Щелчок мышью по любому слову в таблице переводит (аналогично навигации в интернете) на страницу описания данного слова, откуда видны все его грамматические свойства, и морфологический состав. Щелкнув по слову норм|ативн|ый, пользователь попадал на страницу следующего вида:

Слитный текст слова	**нормативный**
Порождающий узел словаря	*норм|ативн|ый*
Грамматический тип	*Прилагательное*
Профиль и основа	**Профиль:** *абажурный* **Основа:** *норм|ативн*

Сегменты слова: ***Н\норм А\ативн ый***	**Корень:***норм* **Суффиксный сегмент:***ативн* **Окончание:***ый*	
Словоизменение:	**Прилаг. именительный падеж муж. ед. число**	**нормативный**
	Прилаг. родительн. падеж ед. ч.	**нормативного**
	Прилаг. дательный падеж мужск. ед.ч.	**нормативному**
	Прилаг. винит. падеж мужск. неодуш. ед.число	**нормативный**
	.....	**.....**

Дальнейшая навигация осуществлялась аналогично. Каждый выбор ссылки перемещал пользователя по воображаемой дуге графа Гиперсловаря от одного узла к другому. Ценность и удобство такого интерфейса трудно переоценить. Однако, содержательная часть словаря, а именно автоматически полученные разбиения, как видно из приведенного выше примера для сегмента "ативн", оставляли желать лучшего. Попыткой достигнуть этого лучшего т.е. лучшей процедуры автоматического морфологического разбора, на основе улучшенных списков морфем, являлся **второй этап** данного проекта, выполнявшийся автором совместно с лингвистами института под научным руководством А.С. Нариньяни осенью 2003-го года. Общей целью, поставленной А.С. Нариньяни, было построение компьютерной реализации универсальной модели русского слова, а в проекции на работу над Гиперсловарем суть его была в следующем. От первого этапа остались списки аффиксных блоков, полученные программным комбинированием законных аффиксов в попытке найти суффиксы в конце или префиксы в начале реального словарного слова. Таким образом получилось много фантомных сочетаний, например неверный префикс СВИС=С+В+ИС в словах типа *свистопляска*. (Хотя приставки С и В не сочетаются, но машине этого априори не известно). Можно было вручную отсеять такие неверные сочетания аффиксов и добавить недостающие, более внимательно проработав материал, чем это мог себе позволить автор на первом этапе. Эта работа, а так же пополнение исходного массива корней была проделана лингвистами РосНИИ ИИ, И.С. Кононенко, О.П. Симоновой, Е.Г. Соколовой, Т.Б. Сосенской. Из числа 530 префиксов "грязного" списка в вычищенном остались только 422. А для суффиксов это число сократилось с 8132 до 1994. Кроме того, автором была соответствующим образом исправлена процедура разбора. В результате качество процедуры автоматического разбиения существенно выиграло. Хотя остались систематические ошибки (с той же приставкой С или приставкой В или с омонимией русских и иностранных корней) но по внешней визуальной оценке автора (а точно посчитать число ошибок в 100-тысячном словнике не представлялось возможным) процент правильно разобранных слов приблизился к 95%. Можно было значительно улучшить этот показатель, например используя частиречную информацию или более глубокий анализ сочетаемости, но в начале 2004-го года проект (точнее его второй этап) был приостановлен по организационным причинам. Вдобавок, к тому моменту обнаружился другой, более простой путь автоматизированного получения точной морфологической информации для Гиперсловаря, а именно: созданная в ИРЯ АН электронная версия словообразовательного словаря А.Н. Тихонова [7] (**СТ**). Это заставило заново пересмотреть всю концепцию работы по Гиперсловарю. Хотя морфемные разбиения сами по себе дают возможность на порядки сократить объем описаний семантической информации по сравнению с прямыми компьютерными аналогами толковых словарей (а ради этой компрессии и затевалась вся работа), словообразовательные связи

вносят в построение семантической сети неизмеримо больший, качественный вклад. Уже сам по себе перевод в форму сетевой базы данных словаря Тихонова устроенного (в отличие от этимологически ориентированного **СКЕ**) по семантическому принципу, закладывал бы основу для семантической сети, позволяя применить при анализе значений слов основные метафоры объектно-ориентированного программирования, в первую очередь - наследование. Словообразовательные связи в подавляющем большинстве случаев можно трактовать и использовать как связи множественного наследования признаков. Множественного наследования потому, что производное слово наследует признаки как минимум от двух источников: от слова основы и от словообразующего форманта. Формантом чаще всего является добавление аффиксного морфа, но в общем случае формант может быть более сложным: 1) не только добавление, но и исключение; 2) могут быть объемлющие форманты; 3) формант может включать чередования, как в корне, так и в аффиксе; 4) некоторые форманты включают перенос ударения.
Помещение всей информации из **СТ** в сетевую базу знаний Абриаля позволяет произвести обобщение информации по формантам, подобно тому, как это было сделано автором с грамматикой по Зализняку [2]. Тогда удалось свести все варианты словоизменения к некоторому компактному набору "профилей" (множеств одинаковых флексий), использование которых позволяло просто и быстро генерировать и анализировать русское словоизменение. Аналогично этому можно поступить и со словообразовательными формантами. В недалекой перспективе словообразовательные форманты и словоизменительные "профили" легко объединяются, т.е. обобщаются до единого механизма, опирающегося на единую базу знаний. Морфологический анализ словоформы можно будет объединить, качественно одним и тем же механизмом переходя от словоизменения к словообразованию. Т.е. словоформа при анализе постепенно раздевается как луковица, применением всех возможных словоизменительных и словообразовательных формантов. Кроме компактности описания такая методология дает возможность анализировать новые, незнакомые, неправильно построенные слова (*хотишь*, не *ложьте* , а они *ложут* , *Чуду-Юду* я и так *победю*).
Этой проблематике, т.е. помещение в уже построенный Гиперсловарь словообразовательной информации из СТ, выделение/обобщение в ней формантных типов, и подготовка к построению единого механизма морфологического анализа, на основе формантов и профилей, посвящается **третий этап** развития проекта Ариадна, проходящий в настоящий момент (весна 2004-го). Содержание и текущее состояние данного этапа раскрывается в этом докладе.

О системе Абриаль 2

Абриаль [4] изначально замышлялся как система для создания и эффективного развития больших "полноязычных" семантических сетей, пригодных в качестве инструментальной среды разработчика систем распознавания для естественных языков. Полное описание архитектурных и функциональных особенностей Абриаля выходит за рамки данной работы. Однако здесь стоит остановиться на нескольких принципиальных моментах даже не самого устройства Абриаля, а подхода, положенного в основу его создания.
Прежде всего - Абриаль изначально при создании был рассчитан на мощность самых современных компьютеров, в частности в следующих моментах:

- База данных Абриаля целиком загружается и работает в памяти компьютера. Реляционные базы данных и SQL и всё подобное – отброшены как устаревшие технологии. Если не будет хватать 0.5Г- 1Г для помещения всей семантической сети по языку - значит, память должна быть еще увеличена.

- Текстовые модули интерфейсной оболочки загружаются из разных источников, в том числе из текстовых файлов, компилируются, отрабатывают и показывают результат своей работы в виде странички на браузере пользователя, и на этом заканчивают свою работу. И всё это происходит за доли секунды каждый раз при выборе пользователем ссылки на экране или по нажатию на экранную кнопку. В таком режиме очень легко отлаживать: после изменения интерфейсного файла в любом текстовом редакторе нажимается "Сохранить", а в окне Абриаля - "Обновить". И сразу же виден результат: измененная работа приложения. Но для этого нужны процессоры, скорость которых измеряется в гигагерцах, которые доступны для массового пользователя всего два-три года.
- В базе данных Абриаля автоматически строятся аналоги индексов по всем теоретическим направлениям доступа от каждого объекта. Это требует как высокой производительности, так и значительных затрат памяти.
- Продукционный механизм ядра системы, организует работу правил и ассоциаций так, что многие ассоциации транслируются на ходу, что для сравнительно простых задач может излишне нагружать процессор, но это же дает возможность динамически в run-time выбирать оптимальные пути поиска решений. Такое решение при статической компиляции было бы невозможно, но с другой стороны уберегает сверхсложные расчеты от проблем с "комбинаторными взрывами".

Таким образом, удалось перебросить на работу машины большинство вспомогательных служебных функций, на которые, как известно, в аналогичных проектах тратится более 90% программистских усилий. А все сохраненные силы и накопленный за десятилетия профессиональной работы программистский потенциал, был направлен на решение центральных проблем, нащупанных в процессе многих итераций: Абриаль - десятая система программирования, созданная автором.
Абриаль написан на языке Си++ в среде Си-Билдера. Исходный код программы весит немногим менее 2Мб.
Программа рассчитана на новое поколение пользователей, для которых интернет - родная среда и навигация через броузер по бесконечной сети - наиболее естественный вид интерфейса с компьютером. Основная метафора организации системы Windows и подобных оконных систем отличается принципиально от метафоры веб-навигации, в следующем:

- В оконной системе пользователь ощущает себя хозяином большого, иерархически организованного множества объектов, разворачивающихся в окна, причем на вершине иерархии находятся объекты рабочего стола пользователя.
- Метафора веб-навигации принципиально отличается тем, что пользователь ощущает себя **частью** бесконечной сети, у которой в принципе нет главной вершины.

От гиперсловаря к семантической сети

Вряд ли нова идея - сделать словообразовательную структуру русского языка основой для семантической сети. Следуя этой идее и добавляя в Гиперсловарь данные о словообразовательных связях из СТ, мы закладываем основу полной семантической сети русского языка. Естественный вопрос - как это реализуется, в частности, по отношению к словам, значения которых не выводится из значения компонент, или выводятся, но не целиком.
Тут используется идея, по-видимому, не очень оригинальная, рассматривать словообразование в рамках парадигмы множественного наследования в Объектно-ориентированном Программировании (ООП), а именно: рассматривать производное слово наследником свойств пары родителей: исходного слова плюс форманта.
Допустим (сначала в понятиях ООП) имеется класс объектов С, наследующий свойства двух других классов: А и В. Тогда, если какое-то свойство не описано специально для класса С, то его реализация ищется в классах родителей (А и В). Возможно, что для класса С вообще не описано никаких собственных свойств, тогда все свойства класса С наследуются им от родителей.

Аналогично этому, в словаре с известной словообразовательной структурой, каждое производное слово имеет как минимум одну родительскую пару: исходное слово и формант (в частном случае формантом может быть добавляемая морфема, например приставка). Аналогично мы будем считать, что если значение слова описано, то это значение перекрывает (полностью или частично) значения, наследуемые от родителей. А если значение слова полностью выводится из исходной порождающей пары, то у него отсутствует собственное семантическое описание.

Например, (все приводимые ниже примеры условны и некорректны с точки зрения лингвистики):

1. Значение слова ПРОБИТЬ=**ПРО**+БИТЬ полностью выводится из исходного слова БИТЬ и форманта (**ПРО**-*). Ему не требуется собственного описания в семантической сети.
2. Значение слова ПРИБОЙ=**ПРИ**+БОЙ имеет собственное описание (морской прибой), которое целиком перекрывает значения, наследуемые от родителей.
3. Слово ПРИБИТЬСЯ=ПРИБИТЬ+**СЯ**, =**ПРИ**+БИТЬСЯ, имеет две порождающие пары, и чаще имеет значение прибиться к компании "бился, бился и прибился" чем прибить себя к стене. Этот случай может рассматриваться как промежуточный между крайностями предыдущих двух пунктов. Т.е. здесь значение наследуется частично и требуется дополняющее семантическое описание, например уточняющее от каких именно родителей, или от какого из возможных значений форманта **ПРИ**-* наследуются значения.

Подробнее о семантической многозначности и омонимии. Многие форманты, особенно аффиксные имеют несколько значений, например, *прибить* может означать и прибить что-то к стене гвоздем или прибить, как убить. Отношение к многозначности формантов непосредственно вытекает из отношения к семантической сети вообще. Семантическая сеть здесь рассматривается не как самоценный артефакт, а как базис приложения предназначенного для смыслового анализа текста. Поэтому априори не предлагается какого-то одного решения, а создается инструмент конструктора семантической сети, в котором можно будет свободно выбирать между спектром возможностей, подобно тому, как программист выбирает те или иные конструкции языка программирования, опираясь на понимание стиля, на интуицию. Здесь конечный результат и критерий для выбора решений - адекватность смыслового анализа текста. Рассмотрим несколько вариантов решения для какого-нибудь простого многозначного форманта, например АТЕЛЬ:

- ПИС-**АТЕЛ-Ь** - деятель, тот кто пишет
- ВЫКЛЮЧ-**АТЕЛ-Ь** - устройство, через которое выключают

Тут возможен выбор как минимум из трех вариантов.

1. Формант **АТЕЛ**-Ь считается однозначным, но не до конца определенным, и в каждое определение порождаемого слова должно быть добавлено дополнение, выбирающее между деятелем и устройством;
2. Создаются два омонимичных форманта **АТЕЛЬ1** - деятель и **АТЕЛЬ2** - устройство. И порождаемые слова связываются со своим вариантом, например, так: ПИС-**АТЕЛЬ1** и ВЫКЛЮЧ-**АТЕЛЬ2**. В этом случае никакого специального описания для порождаемых слов не нужно, все необходимые свойства наследуются от определенного родителя.
3. Формант **АТЕЛ**-Ь считается обобщенным, КТО или ЧТО - не определяется для форманта, а определяется по глагольному корню. Т.е. из свойств, унаследованных от корня должно быть ясно, что данное действие могут делать только люди или же оно воплощено в технике, и так далее…

Третье решение, несомненно, самое качественное, т.е. наиболее гибкое, универсальное и компактное, оно позволяет наиболее эффективным образом управлять развитием

Гиперсловаря, как базиса семантической сети. Но, вместе с тем, этот путь и самый трудоемкий, требующий от конструктора сети наибольшей умелости, как в лингвистическом, так и в программистском плане. Последнее практически сводится к уверенному владению основами объектно-ориентированной методологии, глубокому пониманию концепций наследования, полиморфизма и инкапсуляции.

Стадии построения семантической сети

Итак, намечаются следующие стадии построения гиперсловаря и далее полной семантической сети для русского языка.

1. Сначала из грамматического (СЗ), морфемного (СКЕ) и словообразовательного (СТ) словарей создается Гиперсловарь - сетевая база знаний в среде Абриаля 2. Гиперсловарь становится для будущей семантической сети “скелетом”, который на следующих стадиях обрастает плотью за счет работы конструкторов. Конструкторами на первых стадиях являются лингвистически подкованные математики, затем, постепенно, по мере становления основной структуры, к работе присоединяются лингвисты.
1. Выделяется набор исходных слов, для начала это могут быть исходные слова словообразовательного словаря **СТ**, но не обязательно в точности это множество.
2. Создается набор классов (или классификационных признаков), таких как ”глаголы движения” ”признаки размера”, “шкалы”, “цвета”, “предметы” и эти признаки присваиваются основным исходным словам, причем одному слову присваивается несколько признаков. Выражаясь иначе, одно слово попадает в несколько классов. Этап похож на создание «тезауруса» но с учетом отличия от человеческих тезаурусов отсутствием избыточности: классифицируется не все подряд, а только основные слова. Остальную семантику достроит наследование в словообразовательных и гипернимических связях.
3. На исходных словах или классах вводится структура отношения частное-общее (гипернимии). Впрочем, лучше задавать отношение гипернимии не на словах, а только на классах. Таким образом, этот пункт объединятся с предыдущим. Формально такая система проще, а, следовательно, лучше. Например, рассмотрим связь *селедка-рыба*. *Рыба* здесь гиперним. Будем считать, что для каждого гипернима существует одноименный класс. Например, автоматически появляется связанный со словом *рыба* класс *Рыба*, и слово *селедка* получает классификационный признак, проще говоря, относится в числе прочего к классу *Рыба*. Итак, каждое слово наследует признаки от словообразовательных родителей (основного слова и форманта), и в то же время каждое слово может относиться к классам, через них оно наследует свойства родительских классов. Но не все классы обязательно наследуются, например, если существительное порождается глаголом, то сам класс «глагол» им не наследуется, формант отрицания НЕ-* также специфически модифицирует наследование. Некоторые классы могут быть взаимно несовместимыми и тогда более специальный класс, заслоняет классы более общие.
4. Над множеством классов, которые на этой стадии трактуются как понятия, выстраивается отношения более высокого порядка, в частности часть-целое, антонимии, и другие, вплоть до «*рыбы* ***имеют*** *чешую*», «лошади ***едят*** *сено*» и так далее. Отношение синонимии специально вводить не требуется: любые два слова с одинаковым набором классов считаются синонимами.
5. Наконец, над множеством классов с наследованием, связанных между собой отношениями высокого порядка выстраивается множество грамматических, семантических и прагматических продукционных правил системы Абриаль. Тем самым, мы получаем активную семантическую сеть, используемую как смысловой анализатор русского текста. Принципы функционирования этого анализа в системе Абриаль были изложены автором на Диалоге 2003 [6].

Заключение

В момент подготовки данного доклада первые две из вышеизложенных стадий в основном закончены, а оставшиеся выполняются в экспериментальном режиме, без массового наполнения реальными данными, т.е. производится отладка базовых технологий конструирования в среде системы Абриаль. С материалами по текущему состоянию работ можно ознакомиться на сайте института [1].

Литература

6. http://www.artint.ru/packin/abrial/, http://packin.narod.ru/pro/
7. Зализняк А.А. , Грамматический словарь русского языка, М., Русский язык - 1977
8. Кузнецова А.И., Ефремова Т.Ф. Словарь морфем русского языка. М., Русский язык – 1986
9. Пацкин А.И. Программа ABRIAL - конструктор баз знаний в системе ИНФО-Т. Труды 7-й национально конференции по искусственному интеллекту КИИ-2000. Переславль-Залесский 2000.
10. Пацкин А.И. Гиперсловари на базе системы Абриаль. ДИАЛОГ'2002, Труды межд. семинара. М., 2002.
11. Пацкин А.И. Применение техники управления событиями для анализа текста в системе Абриаль. ДИАЛОГ'2003, Труды межд. семинара. М., 2003.
12. А. Н. Тихонов, Словообразовательный словарь русского языка, Русский язык, 1985