

СЕМАНТИЧЕСКИЕ СВОЙСТВА ИМЕН ДЛЯ ГЕНЕРАЦИИ ЗАПРОСА К БАЗЕ ДАННЫХ

Е. Г. Соколова,
Российский НИИ искусственного интеллекта,
e-mail: sokolova@aha.ru

Ключевые слова: автоматическая обработка текста, генерация текста на естественном языке, семантика существительных, именная группа, однородность

В статье описываются семантические средства генерации запроса на русском языке из исходного формального объектного представления, полученного как результат анализа запроса пользователя системой InBASE. Запрос генерируется в виде распространенной именной группы, соответствующей структуре отношений между объектами и атрибутами формального представления. Рассматриваются семантические свойства, которые необходимо придать понятиям в словаре для выражения отношений между понятиями и для создания естественного порядка в цепочке однородных членов средствами русского языка, а также некоторые дополнительные факторы, влияющие на реализацию ЕЯ запроса.

1. Введение

В статье [Болдасов, Соколова, 2002] мы описали задачу создания генератора запросов на ЕЯ из формального представления. В настоящее время разработана среда DEMLinG [Boldasov, 2003] для создания генераторов, а также создан прототип генератора запросов к кадровой БД QGen, который предполагается использовать как модуль генерации в системе InBASE и расширить на другие тематические области. Задача данной статьи – собрать и описать семантические и языковые (русского языка) средства, которые использовались в прототипе QGen с целью их систематизации и теоретического осмысления в русле генерации ЕЯ высказываний как фундаментального научного направления.

В работах по генерации текстов на ЕЯ в качестве исходного обычно используются семантическое, близкое к глубинно-синтаксическому, представление содержания текста. Оно создается либо пользователем в режиме интерфейса с системой, например, в системе генерации софтверных инструкций AGILE [Соколова, Шаров, 2000], [Kruiff et al. 2000] либо планируются на основе коммуникативной цели высказывания и базы знаний [Макьюин, 1989]. В некоторый момент происходит лексикализация понятий либо как прямая замена понятия на лексему по словарю, либо как вывод лексического воплощения на основе пучка семантических свойств, например, в системе Микрокосмос [Nirenburg, 1987], оба способа используются в AGILE. Для грамматической реализации на конкретном языке создаются модули грамматики конкретного приложения или универсальные с многоязыковыми ресурсами, например, KPML [Bateman, 1997].

Особенность генерации в QGen состоит в том, что мы реализуем средствами ЕЯ непосредственно понятия и отношения между понятиями в предметной области, без использования промежуточного универсального семантического языка или глубинного синтаксиса. Это становится возможным в связи с характером исходного представления запроса, поступающего в QGen.

На вход генератора поступает формальное представление запроса на языке Q – объектно-ориентированном языке представления запросов. Q-представление является XML файлом и имеет древесную структуру. Полнозначные понятия, образующие Q-представление, классифицированы в БД по двум типам: Атрибут и Значение (Атрибута). Кроме того, используются Функторы: max, avg, min, sum, count, и Операторы: like, =, >=, >, <.

В связи с тем, что БД строится без учета возможного выражения ее содержания в виде текста или высказывания на ЕЯ, для того, чтобы проинтерпретировать запрос пользователя, заданный на ЕЯ, в системе InBASE строится

посредническое представление – модель предметной области (МПО), цель которого – приблизить содержание БД к понятиям, которыми обычно пользуется говорящий человек [Sharoff, Zhigalov, 1999]. МПО представляет собой иерархию понятий, представляющих все атрибуты БД а также дополнительные объекты, которые иногда соответствуют таблицам БД. Это те объекты, “носители” атрибутов БД, которые являются целью интереса пользователя. Например, обращаясь к кадровой БД пользователь спрашивает о *сотрудниках*, обращаясь к БД научных статей - об *авторах, городах*, в которых изданы *статьи* или расположены *издательства, журналы, издательства*. Все эти Объекты характеризуются конкретными Атрибутами БД, тоже представленными как понятия в иерархии МПО. Значения Атрибутов при построении Q-представления достигаются через пути, заданные при соответствующих Атрибутах в МПО.

Q-представление является связным выражением содержания запроса, структура которого задана трехчастной структурой подобной SQL выражению (SELECT, FROM, WHERE). Отличие от SQL заключается в том, что составляющая FROM содержит не идентификатор таблицы БД, а имя объекта МПО, обладающего определенными свойствами. Назовем Q-представление метасемантическим, поскольку, хотя оно образовано понятиями неязыковой природы, они рассматриваются в проекции на их выражение в ЕЯ коммуникации. Необходимые для превращения Q-представления в ЕЯ высказывание¹ семантические сведения задаются в словаре QGen’a в процессе настройки ЕЯ интерфейса системы InBASE на конкретную БД [Жигалов, Соколова, 2000]. Они определяются для данного регистра (запрос к БД) и для данного “значения” лексической единицы, обозначающей в точности понятие МПО в его отношении к другим понятиям данной МПО. В данной статье мы рассматриваем генерацию запроса в теоретическом аспекте, опуская детали ее реализации в среде DEMLinG.

Статья имеет следующую структуру: в разделе 2 рассматриваются некоторые свойства Q-представления и его выражения на ЕЯ; в разделе 3 рассматриваются средства русского языка для выражения Атрибутно-Объектных отношений Q-представления; в разделе 4 описывается выражение других метасемантических отношений между понятиями Q-представления. В разделе 5 сказано несколько слов о реализации коммуникативной задачи высказывания при генерации. В разделе 6 ставится задача моделирования порядка слов в сочиненной группе в тексте запроса на ЕЯ. В Заключение перечисляются итоги работы.

2. Q-представление как формализованное содержание ЕЯ запроса

Q-представление является языковнезависимым поскольку оно состоит из понятий МПО и БД и не содержит языковых предикатов, обычно организующих высказывание на ЕЯ. Пример см. в [Болдасов, 2003].

Содержание, выраженное в виде Q-представления, может передаваться средствами ЕЯ в различных формах. Наиболее близкой к Q-представлению ЕЯ формой является грамматическая форма именной группы (ИГ), состоящая из существительных, или словосочетаний с функцией существительного, представляющих понятия МПО. Понятия соединены между собой отношениями, выраженными в запросе на ЕЯ средствами грамматики, например, *Авторы статей в журнале НТИ: “Авторы”* – запрашиваемые Атрибуты – вершина ИГ, *“статей”* – уточнение (или постмодификатор) к вершине, Объект, которому принадлежат запрашиваемые Атрибуты. Отношение между Атрибутом и Объектом (отношение Атрибут-Объект) выражено родительным падежом; *“в журнале НТИ”* – уточнение к Объекту, которое представляет собой пару Атрибут(Значение). Средство русской грамматики, выражающее данное отношение Объект-Атрибут(Значение) – предлог “в” с предложным падежом.

Содержание запроса также может быть выражено в форме предложения с глагольным сказуемым. Мотивом для введения предиката является жанр генерируемого текста – запрос к БД, а не содержание запроса. Предикат вводится в представление волевым решением говорящего либо на основании лексико-семантических отношений (ЛСО) между связываемыми понятиями, в котором они могут состоять как элементы лексико-семантической системы русского языка, либо в качестве лексической реализации коммуникативной задачи высказывания “запрашивание информации”. Например:

- *Кто пишет статьи в журнал НТИ?* Предикат *пишет* выводится из лексико-семантического отношения между понятиями *Кто=автор* и *статья* – результат деятельности автора.
- *Сообщите авторов статей в журнале НТИ.* Здесь выбирается предикат, обозначающий выдачу информации адресату.

¹ В [Болдасов, Соколова, 2002] описывается небольшая чистка Q-представления в процессе планирования, которая связана со спецификой структуры МПО и различиями в обращении системы к БД для выдачи информации в виде таблицы, как это сейчас делается в InBASE, по сравнению с содержанием ЕЯ запроса, характерного для общения людей.

Есть ограниченный тип ситуаций, когда введение предиката необходимо независимо от воли говорящего для того, чтобы запрос был правильным с точки зрения языкового выражения. Этот случай рассматривается в разделе 3.1.2.

Основными отношениями в Q-представлении являются отношения Атрибут-Объект и Объект-Атрибут(Значение). Первое возникает между каждым членом раздела SELECT и членом раздела FROM, второе – между членом раздела FROM и каждым членом раздела WHERE. Эти два отношения и другие отношения между понятиями в Q-представлении назовем метасемантическими. Реализация отношения между выраженными на ЕЯ понятиями Q-представления может происходить как на базе метасемантического отношения, так и на базе ЛСО. Кроме того, отношение может рассматриваться как просто семантическое отношение, если оно установлено на основании семантики связываемых слов.

3. Формы реализации атрибутно-объектных отношений Q-представления на русском языке

Два имени связываются в русском языке с помощью падежных и предложно-падежных форм. Семантика этих средств описывается, в частности, в Синтаксическом словаре Г.А.Золотовой [Золотова, 1988], который будет привлекаться при семантической интерпретации метасемантических отношений. Метасемантическое отношение дублируется лексико-семантическим отношением (ЛСО), существующим между этими понятиями в лексико-семантической системе языка. Таким образом, в исходном Q-представлении понятия Атрибут и Объект связаны двумя отношениями – метасемантическим, в которое они поставлены волей говорящего, и ЛСО, которое мотивировано семантикой одного из связанных понятий или их соотношением.

3.1. Реализация отношения Атрибут-Объект

Метасемантическое отношение Атрибут-Объект между понятиями Q-представления, выраженными субстантивами в ЕЯ запросе, обычно реализуются родительным падежом. Необходимость реализации на базе ЛСО возникает в редких случаях. Для того, чтобы их определить, необходимо рассмотреть источники и типы возможных ЛСО.

3.1.1. *Genitive*

Для беспредложного родительного падежа, которым реализуется метасемантическое отношение Атрибут-Объект, в Синтаксическом словаре Золотовой, примеров почти нет. Большинство примеров иллюстрируют семантические отношения, не свойственные регистру запросов к БД, например, агентивное, предикативное и т.п., что свидетельствует о некоторой искусственности именных запросов, генерируемых на основе Q-представлений, а также о маргинальности жанра запросов к БД. Однако, с точки зрения изучения работы русского языка для выражения мысли, маргинальность жанра запросов не имеет значения.

Для определения ЛСО между двумя понятиями существенным оказывается вопрос об их семантических типах. Выделяется три случая:

1. Понятия разнотипны. Одно из них, Атрибут, является абстрактным понятием, обозначающим признак, тип или параметр предмета, который представлен вторым понятием, например, *имя, возраст, образование и зарплата сотрудника...* или *цвет и марка автомобиля...*. В этих примерах ЛСО мотивировано семантикой абстрактного понятия.
2. Если понятия не являются разнотипными в описанном в п. 1) смысле, то между ними может быть обнаружено ЛСО из тех, которые обычно указываются в лексико-семантических базах данных (ЛСБД), например, WordNet [WordNet]. Это отношения меронимии, например, “часть → целое”, “деталь – устройство”, и другие отношения, например, “обладаемое” → “обладатель” и т.п. Например, *карбюратор автомобиля, количество детей сотрудника*. В [Золотова, 1988, с. 32], отмечено только одно отношение, possessive, причем в примерах используются имена, по своей семантике не свойственные БД: например, *силами своего существа* – слишком абстрактное, *платки девчат, папирус Гарриса* – семантически композиционные: *девчата* – “*провожающие* (или *студенты* и т.п. в зависимости от темы (события, которое описывается) возможной БД) *женского пола в возрасте <15-35 лет>*”, *Гаррис* – *владелец (папируса) по имени Гаррис* .:
3. В том случае, если понятия являются однотипными, например, *начальник* и *сотрудник* оба имеют гипероним *люди*, и между ними не установлено ЛСО типа указанных в п.2, ЛСО может устанавливаться на основании семантики одного из слов (в ЛСБД толкование слова обычно не формализовано). Можно говорить о наличии между понятиями *начальник* и *сотрудник* ЛСО без этикетки, поскольку этикетка лексикализована в слове *начальник*.

Существительным выражаются Функторы sum и count – *сумма* и *количество*. В этом случае они являются косвенным определением Объекта, т.е. выражаются родительным падежом (семантическое отношение “количество” [Золотова, 1989 с. 30]). Sum сочетается в Q-представлении с числовыми Атрибутами типа Зарплата, count – с символическими типа Фамилия, например, (*количество сотрудников*) = *количество Фамилий* в БД. ??

4.2. Реализации отношения Атрибут-Значение

В генерируемом QGen’ом именном запросе Значения включаются в текст запроса в том виде, в котором они находятся в БД, т.е. как неизменяемая этикетка и заключаются в угловые скобки. Соответственно используется грамматическое средство “приложение”, например, *со скоростью <50 км/час>, по имени <Андрей>*. Впоследствии может быть введена минимальная классификация Значений по типам, например, Прилагательное и Неизменяемое. Прилагательные представлены закрытые классом слов, например, Прилагательные Цвета или Пола. К неизменяемым могут быть отнесены Атрибуты, значения которых реализованы в БД в исходной форме существительного или количественным числительным. К таким Атрибутам относятся Имя, Отчество, Фамилия, Скорость_автомобиля и т.д. При генерации Значений, относящихся к указанным типам, угловые скобки будут опускаться, а Значение встраивается в соответствующее место синтаксической структуры: Прилагательное – слева от определяемого существительного, обозначающего атрибут, а Неизменяемое – после определяемого существительного, например, *сотрудник женского пола, автомобиль со скоростью 100 км/час, сотрудник по имени Андрей*.

Мы не планируем генерации существительных с “композиционной” семантикой типа “женщина” вместо *сотрудник* (Объект) *женского* (Значение) *пола* (Атрибут). В QGen, таким образом, преследуется дополнительная цель – отразить в сгенерированных ЕЯ запросах устройство БД.

4.3. Реализации отношения Оператор-Значение

К Значению примыкает Оператор, обозначающий меру запрашиваемого Значения, которая передается фиксированным словом или словосочетанием: *больше, меньше, больше или равный, меньше или равный*. Например, *Фамилии сотрудников с количеством детей больше или равным <3>*.

5. Реализация прагматических аспектов высказывания

Отмеченная выше в разделе 2 возможность лексикализации коммуникативной задачи высказывания глаголами типа *сообщите, выдайте, покажите список* и т.д. поднимает вопрос об экспликации коммуникативной роли исходного представления информации в виде специального узла. Его лексикализация может осуществляться на основе лексико-функциональной сети с проверками в узлах переходов типа грамматической сети в KPM1 в духе функциональной грамматики Хэллiday. В местах выбора сети проводятся проверки, по результатам которых идет дальнейший вывод лексемы. Условия, которые проверяются, не являются семантическими. Они относятся к прагматике и контексту речевого акта, например, *покажите список* нельзя выбрать в условиях устного общения, надо: *перечислите* и т.п.

Это также наводит на мысль об описании речевых функций определенных форм языка, например, ИГ в Именительном падеже. Как свободное высказывание эта группа обозначает заголовок в письменном тексте, представление предмета, который она называет, или запрос информации о предмете (в устной речи, в запросах к БД) или самого предмета (в торговле) в соответствующих условиях непосредственной коммуникации.

К прагматике также относится выбор грамматического числа существительного. Некоторые понятия реализуются именами, имеющими фиксированное грамматическое число, например, *Возраст, Стаж, Пол, Семейное положение*. В других случаях грамматическое число реализованного на ЕЯ понятия зависит от количества упомянутых объектов, например:

Фамилии, имена и должности сотрудников с семейным положением <married>

Домашний телефон, фамилия и имя сотрудника в должности <'president'>

6. Порядок слов

Обычно в системах генерации рассматриваются конструкции, в которых между узлами установлено отношение зависимости, например, предикат с партиципантами, субстантив с определением и т.п. В них порядок слов устанавливается в соответствии с грамматикой данного ЕЯ. В запросе, представленном в виде ИГ, мы рассматриваем цепочки однородных членов, которые формируются в частях SELECT (сочиненная цепочка Атрибутов) и WHERE (сочиненная цепочка Атрибутов со Значениями). Известно, что порядок слов в сочиненных

цепочках не случайный, а подчиняется некоторым принципам. В частности, порядок слов в цепочке приименных определений, выраженных прилагательными, подчиняется строгому семантическому порядку следования. Он формулировался в ранних работах по моделированию языка Мельчуком и Хэллidayем [Halliday 1985], затем подробно исследовался в работах типологического направления. В качестве примера, следующая цепочка семантических позиций определена для существительного (вещь) и его адъективных модификаторов: дейктик – количество – модальность – цвет – качество – вещь. Например, *the two new bright silver rings* (англ.); *эти два новых блестящих серебряных кольца* (рус.); *ces deux bagues d'argent brillantes* (франц.). Принцип семантического упорядочивания прилагательных моделируется в прикладных системах, использующих семантические представления, например, в системе KPMI. В прототипе QGen мы используем специальные семантические признаки, которые определяют порядок следования вершинных элементов в SELECT части и постмодификаторов в WHERE части. Он задается в QGen в терминах семантических признаков, управляющих реализацией отношений Объект-Атрибут(Значение), которые, в свою очередь, отражают различные предложно-падежные формы соединения этих понятий. Таким образом, сочинительная цепочка упорядочивается в конечном счете по синтаксическим формам. Интересен вопрос корреляции этих форм с семантикой в разных языках и о наличии общего синтактико-семантического принципа их упорядочивания. Таким образом, в отличие от прилагательных, здесь можно усматривать три возможных принципа упорядочивания: семантический (закрепленный в грамматике как у прилагательных), синтаксический (по более и менее узкому способу реализации отношения) и прагматический, специфический для конкретной предметной области. При этом на порядок следования элементов также оказывает влияние возможная неоднозначность.

7. Выводы и дальнейшие направления исследований

В статье показано, что:

1. способ реализации не всегда определяется семантическими отношениями единого семантического представления высказывания, которые обычно используются как входные данные для генераторов;
2. Особенность решения задачи генерации в QGen состоит в том, что мы реализуем средствами ЕЯ непосредственно объекты (понятия) и отношения между объектами в объектном представлении ПО без использования промежуточного семантического или глубинно синтаксического представления. В этом случае выделяется по крайней мере три типа отношений: метасемантические – между объектами МПО во входном предметно ориентированном представлении, волей говорящего организующие высказывание (текст) в форме, свойственной данному регистру; ЛСО – между объектами МПО, выраженными лексемами ЕЯ, выражающие постоянные связи понятий в лексико-семантической системе ЕЯ; наконец, семантические отношения, устанавливаемые между двумя понятиями на основе их семантики.
3. Перечислены способы реализации трех видов семантических отношений для понятий кадровой БД. Есть основание утверждать, что Репертуар элементарных единиц русского синтаксиса, описанный Г.А.Золотовой может служить основой для формализации грамматических средств русского языка и их использования в конкретных приложениях. Однако для этого он должен быть расширен и верифицирован на большом корпусе текстов (что в настоящее время становится возможным). Кроме того, подробное исследование по генерации в конкретных приложениях позволит не только использовать эти данные, но и позволит сделать это описание более точным и адекватным, в частности, учесть регистры речи.
4. Необходимо описать речевые функции определенных форм языка, например, ИГ в Именительном падеже как свободное высказывание обозначает Заголовок в письменном тексте, представление предмета, который она называет, или запрос информации о предмете (в устной речи, в запросах к БД) или самого предмета (в торговле) в соответствующих условиях непосредственной коммуникации.
5. Я надеюсь, что задача “собрать” лингвистические явления, участвующие в генерации запроса выполнена в первом приближении. Более подробное рассмотрение требует привлечения дополнительных исследований, например, работ по параметрам С.Семеновой, а также дальнейшего развития генератора QGen в соответствии с представлением о том, что фундаментальное знание извлекается в процессе создания действующих систем. Многие вопросы требуют специального исследования, например, принципы упорядочивания однородных постмодификаторов.

Литература

1. Болдасов М.В., Соколова Е.Г. Генерация содержания запроса пользователя системой InBASE в виде естественно-языкового высказывания // *Труды Международного семинара Диалог'2002 по компьютерной лингвистике, Том 2*, Протвино, Июнь 2002, С. 49-60.

2. Болдасов М.В. Парадигмы генерации ЕЯ текстов в инструментальной среде DEMLinG. // (данный сборник)
3. Жигалов В.А., Соколова Е.Г. InBASE: технология построения ЕЯ интерфейсов к базам данных // *Труды Международного семинара Диалог'2001 по компьютерной лингвистике, Том 2*, Аксаково, Июнь 2000, С. 123-135.
4. Золотова Г.А. Синтаксический словарь. Репертуар элементарных единиц русского синтаксиса. М.: Наука, 1988.
5. Маккьюин К. Дискурсивные стратегии для синтеза текста на естественном языке // НЗЛ. Вып. XXIV. М., 1989. С. 311-356
6. Соколова Е.Г., Шаров С.А. Многоязыковая генерация руководств пользователя на славянских языках // КИИ'2000, Т.1, М.: Изд-во физ-мат. лит-ры, 2000ю С. 409-417.
7. John Bateman; KPML Development Environment – multilingual linguistic resource development and sentence generation. Release 1.1 January 1997 // Institut fuer integrierte Publikations- und Informationssysteme (IPSI), German Centre for Information Technology (GMD), Dolivostr. 15, Darmstadt Germany. URL: <http://www.darmstadt.gmd.de/publish/komet/kpml.html>
8. Kruijff, G-J., Bateman J., Dochev D., Hana J., Kruijff-Korbayova I., Skoumalova H., Sharoff S., Sokolova E., Staykova K., Teich E.; Multilinguality in a Text Generation System for Three Slavic Languages // **Proc. COLING-2000**, Luxemburg, August, 2000.
9. Sharoff, S. & Zhigalov V. Register-domain Separation as a methodology for development of NL Interfaces to Databases // Human-Computer Interaction – INTERACT'99. Angela Sasse and Chris Johnson (eds.) Published by IOS Press, 1999.Mass
10. Ljungberg, A. An information state approach to aggregation in multilingual generation. M.Sc. dissertation, Univ. of Edinburg. 2001. - <http://www.iccs.informatics.ed.ac.uk/~jbos/anna/>
11. Nirenburg, S. Machine translation: theoretical and methodological issues. Cambridge Univ. Press, Cambridge. 1987.
12. WordNet: <http://www.cogsci.princeton.edu/cgi-bin/webwn>
13. Lexical FreeNet: <http://www.lexfn.com>