Семантическая структура пропозиции при извлечении фактов из текстов на русском языке

Азарова И. В.
azic@bsr.spb.ru

Иванов В. Л.
artifex@increatum.com

Овчинникова Е. А.
eo@kiberry.ru

В докладе обсуждается технология обработки текстов на русском языке, разработанная в рамках проекта ИДЕОГРАФ[1]. Из текста извлекается грамматическая, лексико-семантическая, семантическая информация, которая может быть использована для различных практических задач, в частности, для извлечения из текстов фактоидов.

1. Компоненты системы

Отличительной особенностью технологии проекта является использование (а) формализмов AGFL и HPSG для описания грамматической и семантической структуры; (б) компьютерного тезауруса RussNet, фиксирующего семантические связи в лексической системе русского языка; (в) блока семантического анализа, снимающего грамматическую и лексическую неоднозначность, и задающего пропозициональную интерпретацию текста; (г) специальной оболочки ИдеоЛог для интерпретации формальных продукционных представлений разных типов и логического вывода, которая может использоваться на различных компьютерных платформах.

Несмотря на то, что в настоящее время технология реализуется на базе русского языка, в дальнейшем она может быть **расширена** за счет имеющихся грамматик AGFL для **других языков** (английского, голландского, испанского, греческого и др.), тезаурусов wordnet типа для большинства европейских и азиатских языков, которые связаны между собой посредством специального базового набора понятий через особый индекс ILI.

2. Грамматический компонент: RUS4IR

Грамматический компонент представлен двухуровневой контекстно-свободной грамматикой AGFL[2], которая задает полное описание **морфологии и базового синтаксиса русского языка** (Азарова 2003). Результаты анализа лингвистически интерпретируемы. Варианты анализа являются отражением реальной неоднозначности структуры текста.

Особенностью работы грамматического блока является то, что морфологические формы рассматриваются в составе синтаксических конструкций, поэтому высокая неоднозначность анализа слов на морфологическом уровне снижается за счет использования синтаксической информации (например, при использовании падежной формы после предлога или согласовании прилагательного с существительным).

Сочетание синтаксического и морфологического модулей описания позволяет выделять многокомпонентные элементы: аналитические формы глаголов, составные числительные, сложные предлоги и союзы и проч. (*будет слушать, пришел бы, в течение, так как*) даже в тех случаях, когда представлены разрывные составляющие, например, предложные сочетания с местоимениями (*ни у кого, не для кого, друг с другом*).

В морфологическом описании AGFL используется **словарь основ**. Это позволяет избегать нереальных интерпретаций текстовых форм (таких, как аблатив для “салями”), при этом важную роль играют полнота словаря и возможность эффективной обработки больших словарей (Koster 2004).

Для расширения словаря основ в грамматическое описание встроен **блок словообразовательного анализа**, задающий правила образования слов при помощи продуктивных суффиксов и префиксов. В базовый морфологический лексикон включены **частотные основы**. В отличие от регулярных конструкций **исключения** в нашей модели задаются как готовые формы.

Порядок вывода вариантов анализа для фрагмента текста зависит от порядка развертывания правил продукции. Из корпуса современных текстов (Azarova, Sinopalnikova 2004) извлекается информация о частотах реализации каждой синтаксической конструкции. Правила грамматики, описывающие эти конструкции, упорядочиваются по убыванию частот.

---

[1] URL: www.ideograph.ru
[2] URL: http://www.cs.kun.nl/agfl/

В оригинальном формализме AGFL имеется дополнительный ресурс задания предпочтительного варианта анализа penalties[3] («штрафные очки»), которые вычисляются как обратные величины к частотности синтаксической конструкции в корпусе. Для отдельных составляющих в составе порождаемой конструкции «штрафные очки» суммируются, при этом вариант с наименьшим значением является наилучшим.

Автономный **результат грамматического анализа** текста представляет собой выявление

(а) набора лемм для текстового фрагмента;
(б) категориальных (частеречных) показателей леммы;
(в) набора значений грамматических категорий, ассоциированных с текстовым фрагментом;
(г) множества пар слов, связанных синтаксическим отношением главное-зависимое и организованных в одну структуру – дерево зависимостей;
(д) обобщенно-функциональную разметку фрагментов текста (подструктур дерева) в терминах классификации групп слов (например, именное словосочетание, количественное словосочетание, предложно-падежное словосочетание и проч.) и ядерных структур предложений (например, предикативный центр, именной координируемый предикат, прямое дополнение, прямое дополнение под отрицанием и проч.)

3. Семантический компонент

Семантический компонент системы ИДЕОГРАФ состоит их двух тесно взаимодействующих модулей: лексико-семантического, обеспечивающего связь с компьютерным тезаурусом, и синтактико-семантического, при помощи которого из текста выделяются элементарные пропозиции.

3.1. Лексико-семантический компонент: RussNet тезаурус

Лексико-семантическая информация задается при помощи компьютерного тезауруса RussNet, который построен в соответствии с рядом основополагающих принципов создания wordnet-словарей[4]. Элементарной единицей тезауруса является набор синонимичных лексем – **синсет**; слова могут входить в несколько синсетов в зависимости от числа их значений. Допускается включение в синсет устойчивых словосочетаний. Синсеты связаны родовидовыми отношениями в набор **семантических деревьев** (например, *время, совокупность, еда, растения, животные, человек, двигаться, говорить, думать* и проч.) На узлах семантических деревьев определены и другие отношения, например: часть-целое, антонимия, каузация и проч. Синсеты национального тезауруса соотносятся с Межъязыковым лингвистическим индексом (ILI).

Общие для wordnet-лексиконов характеристики RussNet: словарь опирается на сбалансированный **корпус** современных текстов; ядерная структура тезауруса задается примерно двумя тысячами наиболее частотных слов (существительных, глаголов, прилагательных, наречий); разные значения некоторого слова в тезаурусе пронумерованы в соответствии с частотностью их употребления в корпусе текстов; в тезаурусе представлена нетерминологическая лексика.

Расширение свойств русского варианта wordnet-словаря включает введение доминанты синсета (наиболее частотного нейтрального способа выражения лексического значения) и контекстных маркеров значений в виде рамок валентностей, а также исключение окказиональных (редко встречающихся) значений из тезаурусного описания.

Лексический компонент получает входные данные в виде выявленных в тексте лемм и значений их морфологических категорий, при автономном использовании он выдает

(а) набор идентификаторов синсетов, в которые входят леммы с указанной частеречной принадлежностью;
(б) набор вершин семантических деревьев RussNet, в которые входят данные синсеты;
(в) набор имеющихся рамок валентностей для значений.

3.1.1. Снятие неоднозначности при помощи валентных рамок

Помимо структуры лексических значений, в тезаурусе RussNet содержится информация о контекстах употребления слов, которая используется для разграничения значений при построении тезауруса. Для части синсетов определены рамки валентностей, представляющие собой описание зависимых элементов контекста для признаковых слов: глаголов, прилагательных, а также их дериватов (активная валентность), и главных элементов для зависимых: наречий и существительных (пассивная валентность). Эта информация обеспечивает связь между грамматическим и лексическим компонентами и позволяет снимать неоднозначность на обоих уровнях.

Рамка валентностей состоит из перечня позиций элементов контекста. Каждая валентная позиция имеет характеристику **обязательности/факультативности**. В первом случае позиция регулярно (более чем в 67%

[3] http://www.cs.ru.nl/agfl/papers/manual.pdf

[4] Подробно методика построения тезауруса описана в статьях (Азарова и др. 2004; Azarova et al. 2004) и на сайте Санкт-Петербургского университета: http://www.phil.pu.ru/depts/12/RN/

случаев) реализуется в контекстах, причем, даже если в рассматриваемом предложении такой элемент отсутствует, он задан для некоторого предшествующего слова или подразумевается в более общем контексте. Факультативная валентность представлена не так последовательно (более чем в 35% контекстов), хотя является существенной для разграничения значений. Элементы контекстного окружения, появляющиеся с меньшей частотой, считаются окказиональными (незначимыми). Например, для глагола *направиться* в значении 'двигаться в каком-либо направлении' активная рамка валентностей включает две обязательные позиции (субъект и направление движения).

Грамматическая характеристика валентной позиции задает чаще всего предложно-падежную форму, в которой она регулярно выступает в контекстах. Например, первая валентность глагола *направиться* выражается формой именительного падежа, а вторая – либо предлогом "в" в сочетании с винительным, либо предлогом "к" в сочетании с дательным падежом имени. Указанные характеристики покрывают 97% контекстов для первой валентности и 71% для второй, при этом другие способы текстового выражения пункта назначения для глагола движения встречаются существенно реже. Из диапазона вариантов грамматического выражения валентности отбирается лишь устойчивая, ядерная часть.

Некоторые грамматические формы имеют стандартный способ преобразования, например, изменение формы прямого дополнения переходных глаголов под общим отрицанием на генитив (*создавать помехи – не создавать помех*). Рамка переходного глагола не будет включать эту контекстно-обусловленную форму. Такие случаи разрешаются при помощи общих семантических правил трансформации валентных рамок (например, Vt Nacc => neg Vt Ngen).

Семантическая квалификация валентностей осуществляется посредством отсылок к семантическим деревьям тезауруса RussNet. Например, первая валентная позиция глагола *направиться* задает отсылку к дереву “человек”. В других случаях семантическая квалификация может указывать на часть семантического дерева или на отдельный синсет. Регулярно, отсылка дается на специальные наборы семантических деревьев RussNet, например: “одушевленные” (“человек” и “животные”), “предметы” (“естественные объекты”, “артефакты”, “вещество” и т.д.), “сущности” (“одушевленные” и “предметы”).

Для **снятия неоднозначности** используются выходные данные грамматического и лексического компонентов (описанные выше): набор деревьев зависимостей с размеченными узлами и соответствующий деревьям набор лемм с выявленными синсетами RussNet, которые в свою очередь отнесены к определенным семантическим деревьям и снабжены рамками валентностей. Далее происходит сопоставление грамматических и семантических параметров активных валентных рамок с имеющимися грамматическими и семантическими атрибутами текстовых фрагментов. Если вариант анализа фрагмента удовлетворяет полностью валентной рамке, то он считается верифицированным. Если несколько вариантов анализа будут верифицированы, то неоднозначность не будет снята полностью[5].

Рассмотрим частичное снятие неоднозначности на примере фразы “я был знаком с тобой”. На этапе грамматического анализа выделяются две интерпретации: форма *знаком* может быть (а) краткой формой прилагательного *знакомый* и (б) аблативом существительного *знак*. В RussNet прилагательному *знакомый* в предикативной функции соответствуют два синсета с факультативной валентностью на объект, которая грамматически оформлена сочетанием предлога "с" и творительного падежа. Семантическая квалификация объекта в рамке первого синсета указывает на существительное из семантического дерева “человек” ('состоять в знакомстве с кем-либо'), в другом – на группы деревьев “предмет” ('известный, встречавшийся прежде'). Личное местоимение 2-го лица (*тобой*) позволяет выбрать первый из синсетов и первый из вариантов синтаксического анализа. Имеющаяся валентность существительного *знак* (*знак остановки, приоритета*) не подкрепляется контекстом.

Синтаксическая семантика: пропозициональная структура

Модуль синтаксической семантики представляет собой набор семантических интерпретаций, каждая из которых соответствует определенному правилу грамматики. Выходом синтактико-семантического модуля является набор семантических структур определенного типа (объектов, пропозиций и др.). Результатом анализа словосочетания является пропозиция, представляющая в нашей модели описание элементарного факта (фактоида). Выделение таких фактоидов их текста является целью семантического анализа в системе ИДЕОГРАФ.

Пропозициональная структура включает в себя: (а) ссылку на синсет RussNet, соответствующий предикативному центру предложения, например, {*бегать*$_1$} для предложения *Животные быстро бегают по клетке*; (б) список аргументов (функциональных позиций) пропозиции через ссылки на синсеты: {*животные*$_1$}; и (в) список признаков пропозиции: время, место, качество ситуации и др.: {быстро$_1$}, {клетка$_3$},

Частично разметка функциональных позиций задана в валентных рамках RussNet. Традиционно для обозначения функций используется фиксированный набор семантических ролей, однако разграничение ролей (например, объект и результат, адресат и пациенс) не всегда носит объективный характер, поэтому в нашей схе-

[5] Неоднозначность не будет снята также, если у слов нет валентных рамок в тезаурусном описании или если текстовый фрагмент не содержит явных контекстных маркеров.

ме функциональные характеристики различают валентности в рамках некоторого семантического дерева RussNet. Аналогичный подход представлен в проекте FrameNet[6], однако в нашей концепции мы не стремимся "зарезервировать" все значимые позиции для семантического дерева целиком и дать им индивидуальные названия, а допускаем возможное уточнение функций для синсетов с наиболее конкретным содержанием (листьев семантического дерева) или даже для отдельных элементов синсета.

Ядерная группа функциональных значений в структуре пропозиции включает субъект и несколько объектов. Функция ***субъект*** не обозначает активный элемент ситуации ("агенс"), а только указывает на то, что для пропозиций данного семантического дерева такая функция существенна. Функция субъекта пропозиции может быть представлена списком. Например, для фразы *В финале дрались Разяпов и кубинец Мантилья* (а также *В финале Разяпов дрался с кубинцем Мантильей*) позиция субъекта представлена списком, состоящим из двух элементов: [*Разяпов; Мантилья*].

Различные объектные функции приписываются валентным позициям не столько исходя из каких-либо скрытых смыслов, сколько опираясь на разграничение нескольких позиций в рамках пропозиции. Например, для глагола *бросать* в значении 'взмахом заставлять лететь что-либо, находящееся в руках' валентная рамка включает только одну позицию ***объект$_1$*** вне зависимости от того, как этот объект будет выражен в поверхностной текстовой структуре: *бросать камни* или *камнями*.

Несколько объектных позиций вводятся в тех случаях, когда они четко противопоставлены в структуре пропозиции. Например, глагол *забить* в значении 'ударами загнать что-либо куда-либо', исходя из определения, имеет две объектные позиции: собственно объект, к которому приложено действие, и другой объект, на который воздействие происходит посредством первого. Если ролевая квалификация первого объекта очевидна, то для второго она выглядит не совсем ясной. В нашей концепции эти позиции будут различаться как ***объект$_1$*** и ***объект$_2$***. Кроме того, анализ контекстов употребления глагола в корпусе показывает, что помимо указанных объектов, может встречаться еще один (*забить гильзы ватой*), т.е. ***объект$_3$.*** Таким образом, нумерация объектов происходит в соответствии с частотностью употребления этих позиций в корпусе для синсетов из некоторого семантического дерева, причем некоторые из объектных позиций доминируют для большей части синсетов данного дерева.

Семантический компонент содержит набор правил семантической интерпретации для синтаксических конструкций грамматического компонента. Известно, что количество потенциальных интерпретаций конструкции может быть очень большим, однако, мы используем данные корпуса текстов, выделяя и на этом уровне стереотипные, частотные случаи. Ниже мы рассмотрим примеры простейших правил.

На данный момент нами проработаны семантические интерпретации для ряда конструкций именного словосочетания. В частности, если в качестве главного слова выступает дериват признакового слова (отглагольное существительное), то при регулярном соотношении основ существительное анализируется в словообразовательном блоке через соотношение с глагольным синсетом, при этом суффикс задает определенный тип из набора семантических отношений тезауруса RussNet: для существительного *создание* будет установлено отношение der_transposition_action (деривационная транспозиция) с синсетом {*создавать$_1$*}, а для *создатель* – der_agent (деривационный деятель) для этого же синсета.

Семантические правила включают характеристику семантических отношений или семантических деревьев RussNet, например, для присоединения одиночного генитива к существительному наиболее стереотипные правила интерпретации имеют следующий вид:

(1) N $^{\text{der\_transposition\_action}}$ + Ngen $^{\text{ents-id}}$ => P* [object$_1$:Ngen]

Первое правило описывает примеры типа *создание традиции, вскармливание детенышей*, главным словом конструкции является "деривационный транспозит" (N $^{\text{der\_transposition\_action}}$), семантическая квалификация зависимого слова указывает на группу семантических деревьев "сущность"(N $^{\text{ents-id}}$), результатом интерпретации является пропозиция (P*), ядро которой – глагольный синсет, с которым связан транспозит, позиция субъекта не заполнена, позицию первого объекта занимает зависимое существительное.

(2) N $^{\text{der\_agent}}$ + Ngen $^{\text{ents-id}}$ => P* [subject: Var] [object$_1$: Ngen]

Второе правило описывает примеры типа *создатель традиции, проповедник реинкарнации*, главным словом конструкции является "деривационный деятель" (N $^{\text{der\_agent}}$), семантическая квалификация зависимого – "сущность" (N $^{\text{ents-id}}$), результат – пропозиция (P*), ядром которой является глагольный синсет, с которым связан деятель, субъект обозначен как переменная (Var), т.е. референтно-связанный текстовый элемент, позицию первого объекта занимает зависимое существительное.

Синтактико-семантический компонент тесно взаимодействует с лексико-семантическим. Семантические правила задают общую интерпретацию синтаксических конструкций, которая может быть "переписана", если для лексических элементов, входящих в словосочетание, определена рамка валентности, задающая иное пропозициональное представление. Например, описанная выше конструкция с генитивом имеет также субъект-

[6] www: http://www.icsi.berkeley.edu/~framenet/

ную интерпретацию зависимого элемента, если позиция занята синсетом из дерева "человек". Это приводит к тому, что словосочетание *прием посла* будет иметь две семантические интерпретации [object$_1$: {*посол$_1$*}] и [subject: {*посол$_1$*}]. Однако для дериватов глаголов без объектной валентности типа *страдания людей* и для дериватов глаголов с объектной валентностью и семантической квалификацией "предмет" типа *нарушение таксиста* сохранится только субъектная трактовка.

Результаты

Таким образом, интерпретация текстового фрагмента в терминах пропозициональной структуры позволяет очистить анализ от незначимых различий в поверхностной текстовой форме выражения смысла. Соотнесение структуры пропозиции с семантическими деревьями RussNet дает возможность представить смысл в обобщенном или, наоборот, частном виде, содержащем максимальное количество деталей пропозиции. Обобщенный вид пропозиций в дальнейшем предполагается использовать для описания коммуникативной перспективы – постепенного усложнения нескольких микротем, которые составляют сущность информационного содержания текстового документа.

Литература

Azarova I., Sinopalnikova A. *Adjectives in Russnet* // International WordNet Conference, GWC 2004 Brno, Czech Republic, January 20-23, 2004. P. 251-259.

Baker, Collin F., Fillmore, Charles J., and Lowe, John B. (1998): *The Berkeley FrameNet project.* In Proceedings of the COLING-ACL, Montreal, Canada.

Koster C.H.A. *Transducing Text to Multiword Units* // Workshop on MultiWord Units MEMURA at the fourth International Conference on Language Resources and Evaluation, LREC-2004. Lisbon, Portugal, May 2004.

Азарова И.В. *Морфологическая разметка текстов на русском языке с использованием формальной грамматики AGFL* // Компьютерная лингвистика и интеллектуальные технологии. Труды Международной конференции Диалог'2003 (Протвино, 11-16 июня 2003 г .) М., 2003. С . 51-55.

Азарова И.В., Митрофанова О.А., Синопальникова А.А. *Компьютерный тезаурус русского языка типа WordNet* // Компьютерная лингвистика и интеллектуальные технологии. Труды Международной конференции Диалог'2003 (Протвино, 11-16 июня 2003 г .) М., 2003. С . 43-50.

Азарова И.В., Секликов Ю. В., Иванов В. Л. *Интерпретация текстовых документов с использованием формальной грамматики AGFL и компьютерного тезауруса RussNet* // Компьютерная лингвистика и интеллектуальные технологии. Труды Международной конференции Диалог'2004 (“Верхневолжский”, 2-7 июня 2004 г .) М., 2004. С. 1-6.

Азарова И.В., Синопальникова А.А., Яворская М.В. *Принципы построения wordnet-тезауруса RussNet* // Компьютерная лингвистика и интеллектуальные технологии. Труды Международной конференции Диалог'2004 (“Верхневолжский”, 2-7 июня 2004 г .) М., 2004. С. 542-547.