5. Висновок

ГІС - технології надають потужні функціональні можливості аналізу та оцінки просторових даних, чим піднімають на новий, більш високий рівень екологічний моніторинг, що дозволяє підвищити якість та швидкість виконання робіт при вирішенні різного роду задач, прийняття рішень.

Література

1. Україна. Екологічні проблеми атмосферного повітря. Карта Масштабу 1:200 000 . – Київ, ВКФ ТЗ ЗС України – 2000.
2. КНД 211.1.1.106-2002 про КЗ.

Пропонується метод аналізу словосполучень природної мови, котрий основано на формальній граматиці та лінгвістичній базі знань, представленої семантичним словником та онтологією семантичних примітивів

Ключові слова: ПМ, граматика, семантичні функції

Предлагается метод анализа словосочетаний естественного языка, основанный на формальной грамматике и лингвистической базе знаний, представленной семантическим словарем и онтологией семантических примитивов

Ключевые слова: ЕЯ, грамматика, семантические функции

A method for the analysis of natural language phrases, based on a formal grammar and linguistic knowledge base provided by the semantic dictionary and ontology of semantic primitives

Key words: NL, grammar, semantic functions

УДК 001.891:65.011.56

СИНТАКСИКО-СЕМАНТИЧЕСКИЙ АНАЛИЗ ЕСТЕСТВЕННО-ЯЗЫКОВЫХ КОНСТРУКЦИЙ

Г.Ф. Дюбко
Кандидат технических наук, профессор*
Контактный тел.: 702-14-46
E-mail: prof_dubko@yandex.ru

В.И. Омельченко*
Контактный тел.: 063-448-01-51
E-mail: omelya_88@mail.ru

Д.В. Преснякова
Аспирант*
Контактный тел.: 702-14-46
E-mail: Darya_Presnyakov@mail.ru
Кафедра «Программное обеспечение электронных вычислительных машин»
Харьковский национальный университет радиоэлектроники
ул. Серповая, 2, г. Харьков, 61166

1. Ведение

Развитие WEB-индустрии, электронного документооборота, задач искусственного интеллекта, в частности машинного перевода и распознавания речи, стимулировало интерес к автоматическому анализу естественного языка (ЕЯ).

Автоматический анализ ЕЯ является многоаспектной задачей. Один из аспектов анализа – это анализ текстов ЕЯ, результатом которого служит синтаксико-семантическая структура текста. В процессе анализа нужно использовать не только синтаксическую, но и семантическую информацию об анализируемом тексте. Такой подход потребовал создания моделей смысла, одной из которых является модель лексической семантики.

В реальной языковой практике человека лексическая семантика отражена в толковых словарях. На-

пример, в русском языке можно сослаться на словарь [1], который используется для иллюстрации положений, предлагаемых в данной работе. Информацию, содержащуюся в толковом словаре необходимо формализовать, чтобы ее можно было использовать в анализе. Формализация толкового словаря может быть выполнена на основе семантических функций [2], формул лямбда-исчисления, формул логики предикатов.

2. Постановка задачи

Синтаксический анализ осуществляется на основе вывода анализируемой цепочки в формальной грамматике. В основном рассматриваются синтаксико-свободные грамматики. Синтаксическая структура анализируемых текстов описывается продукциями грамматики. В основе вывода лежат две стратегии: свертка и развертка. Свертка предполагает замену правых частей продукции левыми, а развертка – замену левых частей правыми, до получения аксиомы из исходной строки при свертке, и получения исходной строки из аксиомы при развертке. В большинстве случаев процесс вывода является недетерминированным, что приводит к перебору вариантов и неэффективности анализатора. Специальные типы грамматик, такие как LL(K) и грамматики предшествования, позволяют осуществить бесперебороный анализ. Выбор продукций на каждом шаге анализа производится с помощью специально составленных управляющих таблиц.

Формальные языки отличаются тем свойством, что каждой синтаксической структуре поставлена в соответствие единственная семантическая структура, что можно учесть при составлении продукций грамматики. В естественном языке одной и той же синтаксической структуре могут соответствовать разные смыслы. Поэтому механическое перенесение методов формальной грамматики на ЕЯ приводит к огромному числу продукций. Если же учесть, что единицы ЕЯ многозначны, а их значение (смысл) определяется только в контексте, то описание этих единиц продукциями контекстно-свободной грамматики становится весьма проблематичным. Однако отказываться от методов формальных грамматик в ЕЯ не стоит. Поэтому в работе ставится задача разработки синтаксического анализатора, работающего по стратегии свертки методом «перенос-свертка». Грамматика для такого анализатора имеет специальную форму, которая так же подлежит разработке, и имеет информацию о семантике.

Анализатор должен работать с простыми повествовательными предложениями русского языка, главным словом которых является глагол, обозначающий действие. Внутренние словосочетания предложения составляют периферию. Периферия должна сочетаться по смыслу с глаголом. Информация о смысле содержится в семантическом словаре. Для вычисления семантических примитивов используются семантические функции и онтология семантических примитивов.

Входом анализатора служит предложение, где каждое слово разобрано морфологически и заменено с помощью специальной структуры.

Выходом анализатора является семантическая функция, которая представляет лексическую семантику разобранного предложения.

3. Семантические функции и семантические примитивы

Рассматриваются семантические функции двух типов: типа V и типа F, которые отражают лексическую семантику одиночного слова и словосочетания соответственно. $V_i(t)$ обозначает i словарную статью слова t в семантическом словаре, т.е. лексическое значение, в котором употребляется слово t.

Для словосочетаний функция имеет вид F_имя(X_1,...,X_2), где имя – идентификатор словосочетания, X_1,...,X_n – аргументы функции.

Приведем некоторые из наиболее встречающихся словосочетаний, их семантические функции и примеры языковых форм, соответствующие этим словосочетаниям.

Словосочетание «существительное-существительное в родительном падеже» (генитив). Ему соответствует семантическая функция F_генитив(X_1,X_2), где X_1,X_2 либо семантические функции V, либо – F. Например, словосочетание «портфель профессора» имеет семантическую функцию F_генетив(V(портфель),V(профессор)). Словосочетание «портфель профессора математики» имеет семантическую функцию F_генитив(V(портфель),F_генитив(V(профессор), V(математики))).

Словосочетание «прилагательное-существительное». Ему соответствует семантическая функция F_прилагательное(X_1,X_2), где X_1 – существительное или семантические функции с ним, X_2 – функция V от прилагательного. Например, словосочетание «черный кожаный портфель» – F_прилагательное(F_прилагательное((V(портфель),V(кожаный)),V(черный)). Словосочетание «черный портфель профессора» имеет функцию F_прилагательное(F_генитив(V(портфель),V(профессор)),V(черный)).

Словосочетание «предлог-существительное». Ему соответствует семантическая функция F_предлог_1(X_1,X_2), где X_1 – информация о предлоге, X_2 – информация об объекте. Например, словосочетание «на столе» имеет функцию F_предлог_1(V(на),V(столе)).

Словосочетание «существительное-предлог-существительное». Ему соответствует семантическая функция F_предлог_1(X_1,X_2,X_3), где X_1 – информация о предлоге, X_2 – информация о первом существительном, X_3 – информация о втором существительном. Например, словосочетанию «портфель на столе» соответствует функция F_предлог_2(V(на),V(портфель),V(стол)), а словосочетанию «черный портфель на письменном столе» соответствует функция F_предлог_2 (V(на), F_прилагательное (V(портфель), V(черный)), F_прилагательное (V(стол), V(письменный))).

Простое повествовательное предложение, построенное на основе глагола, обозначающего действие, можно считать особым видом словосочетания. Совокупность лексических значений элементов предложения определяет осмысленность предложения. Семантическая функция сопоставленная предложению, имеет вид F_предложения(X_1,...,X_n), где X_1 – информация о главном слове предложения (глаголе), X_2,...,X_n – периферия предложения. Каждый аргумент семантической функции предложения играет свою роль. Если упорядочить аргументы по ролям, X_2 играет роль объекта, выполняющего действие, X_3 – обстоятельство (времени, места и т.д.). Аргументы можно не

упорядочивать, а приписывать им роль в явном виде. Например, предложение «Портфель профессора стоит на письменном столе» имеет семантическую функцию

F_предложения (V1(стоять),
F_генитив (V1(портфель), V1(профессор)),
F_предлог_1 (V1(на),
F_прилагательное (V1(стол), V1(письменный))).

Слова ЕЯ можно рассматривать как семантические примитивы, названия множеств, имеющих определенный состав элементов. Семантические примитивы можно трактовать как категории. Категории можно организовать как онтологию, иерархию, где одни категории становятся подкатегориями других. Верхним онтологии семантических примитивов является ОБЪЕКТ, АБСТРАКЦИЯ, ДЕЙСТВИЕ, ОТНОШЕНИЕ. Подкатегории онтологии можно построить, опираясь на толковый словарь, превращенный в семантический, где каждая словарная статья выражена семантической функцией. Семантическая функция должна быть построена так, чтобы ее главным словом являлся семантический примитив, задающий категорию, описанную семантической функцией. Например, слово «портфель» в толковом словаре описано как «сумка», «сумка» в свою очередь, описано как «хранилище», а «хранилище» – как «объект». Вырисовывается естественная иерархия ОБЪЕКТ-ХРАНИЛИЩЕ-СУМКА-ПОРТФЕЛЬ. Название категории (значение семантического примитива) можно вычислить из семантической функции, задающей семантику слова. Надкатегории и подкатегории вычисляются из онтологии. Семантические примитивы – надкатегории задают более общий смысл, подкатегории конкретизируют смысл.

4. Анализ

Анализ предложения разбивается на два этапа. На первом этапе происходит выделение глагола и словосочетаний, составляющих периферию предложения. В процессе выделения происходит замена словосочетаний на семантические функции. Реализуется первый этап с помощью метода «перенос-свертка».

На втором этапе происходит контекстная оценка словосочетаний в рамках словарных статей выделенного глагола. Реализация второго этапа осуществляется на базе семантического словаря, где контекст представлен лямбда выражением. Результатом работы второго этапа является словарная статья глагола, которая определяет лексическую семантику предложения, и семантическая функция, соответствующая разобранному предложению

Опишем метод «перенос-свертка», который базируется на выводе сверткой в контекстно-свободной грамматике. КС-грамматика – это четверка $G=(V_T,V_N,P,S)$, где V_T – терминальный алфавит, V_N – нетерминальный алфавит, P – множество продукций типа $A \to a$ при $A \in V_N$, $a \in (V_T \cup V_N)$, S – аксиома. Для более гибкого описания синтаксической структуры и введения в нее семантических элементов терминальным и нетерминальным символам приписываются атрибуты, обладающие определенными значениями.

Если X-символ алфавита грамматики, то X.a1 обозначает атрибут этого символа, а запись X.a1=«b»означает, что значение атрибута равно «b».

Метод использующий три структуры данных: ВХОД, где записана последовательность терминальных символов; СТЭК – стек, где могут быть терминальные и нетерминальные символы; УТ – таблица, которая в зависимости от текущего контекста ВХОДА и СТЭКА выполняет операцию переноса или свертки. Операция переноса состоит в записи текущего терминала в СТЭК и сдвижки на один символ по ВХОДУ. Операция свертки состоит в замене правой части продукции, находящейся в верхушке стека, на левую. Работа метода осуществляется по шагам. На каждом шаге проверяется контекст и выполняется соответствующая этому контексту операция.

Для ЕЯ метод «перенос-свертка» требует на входе последовательность слов (предложение), разобранную морфологически. В этом случае каждое слово трактуется как терминальный символ t , имеющий атрибуты. (атрибутами являются часть_речи, падеж, число, род время и т.д.). Например, слово «портфелями» имеет следующие значения атрибутов: t.часть речи_=«существительное», t.род=«муж», t.число=«множ», t.падеж=«творительный».

Грамматика, используемая в процессе анализа, должна описывать словосочетания периферии предложения. Ее терминальный алфавит состоит из терминального символа t. Нетерминальный алфавит состоит из символов, являющихся именами семантических функций. Например, <F_генитив>, <F_прилагательное> есть нетерминалы грамматики. Все символы являются атрибутными и могут иметь индексы для манипуляций с атрибутами.

Продукции грамматики состоят из трех компонент. Первая компонента – это структура продукции, т.е. состав символов и их конкатенация. Вторая компонента – условия применения продукции, сформулированные на знаниях атрибутов. Третья компонента – процедура, определяющая результат применения продукции. Например, продукция, определяющая словосочетание, которому соответствует «генитив» имеет вид $<F\_генитив> \to (t_1,t_2)$. Это структурная компонента. Вторая компонента состоит из условий: t_1.часть речи_=«существительное», t_2.часть речи_=«существительное», t_2.падеж=«родительный». Процедура из третьей компоненты должна: заменять в стэке t_1 t_2 на <F_генитив>; приписывать атрибутам нетерминала значения атрибутов главного слова, т.е. t_1 ; формировать семантическую функцию F_генитив($V(t_1)$,$V(t_2)$), запоминая ее в атрибуте <F_генитив>.семант_функция.

Продукции грамматики строятся на основании языкового экспериментального материала, т.е. на основе анализа словосочетаний, которые могут быть периферией.

Управляющая таблица анализатора должна давать возможность выполнять следующие операции в ректном порядке: перенос, свертка, определение основы для свертки, формирование ω_i, α и очистка стэка. Строится управляющая таблица так же как и продукции, на базе анализа экспериментального материала. В управляющей таблице сформированы условия, характеризирующие контекст входа и стэка на значениях атрибутов, и этим условиям сопоставляет действия. Например, пусть в стэке находится единый терминальный символ t , а на входе текущим символом является глагол. В этом случае необходимо выполнить

свертку для основы размером в один символ. Свертка выполняется в соответствии с продукцией <V> → t, в результате чего семантическая функция сформируется в <V>.семант_функция. Далее формируется ω_i <V>. семант_функция и очищается стэк.

Второй этап анализа основывается на семантическом словаре, где каждая словарная статья глагола связана с λ-формулой. λ-формула определяет смысловой контекст глагола и имеет следующий вид:

$$\lambda X_1 ... \lambda X_n (P_1(X_1) \wedge ... \wedge P_n(X_n) \wedge Q_1(X_1 ... X_n) \wedge ... \wedge Q_m(X_1 ... X_n)) @\omega_1 ... @\omega_n,$$

где $P_i(X_i)$, $Q_j(X_1 \ldots X_n)$ $1 \le i \le n$, $1 \le j \le m$ и P_i – предикатная константа, совпадающая с семантическим примитивом, Q_j – предикат, обозначающий условие на переменных $X_1 ... X_n$. $\omega_1,...,\omega_n$ – словосочетания, выделенные на превом этапе анализа.

Процедуры и данные, которые они обрабатывают приведены на рис. 1.

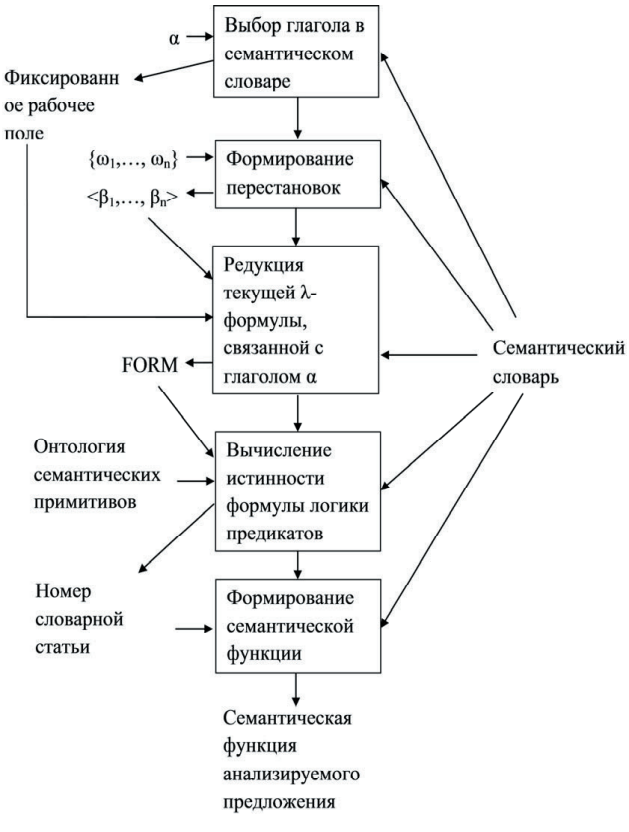

Рис. 1. Схема вычисления словарной статьи и ее семантической функции

Приведем пример λ-формулы, которая определяет контекст при вычислении словарной статьи глагола на втором этапе анализа. Пусть рассматривается предложение «Поезд идет в Киев», где «идет» имеет смысл «объект перемещающийся в какое-то место». В этом случае λ-словарной статьи (перемещение) имеет вид:

$$\lambda X_1.\lambda X_2.\lambda X_3.\lambda X_4 (\text{ОБЪЕКТ}(X_1) \wedge \text{F\_предлог\_1}(X_2) \wedge \text{в}(X_3) \wedge \text{ПЕРЕМЕЩЕНИЕ}(X_3) \wedge \text{ОБЪЕКТ}(X_4) \wedge X_4.\text{падеж}=\text{«именительный»} \wedge X_4.\text{падеж}=\text{«винительный»})$$

$$@\omega_1@\omega_2@\omega_3@\omega_4$$

5. Выводы

Предложенный метод анализа позволяет формировать разбор предложений с учетом семантических аспектов, что значительно повысит качество анализа.

Литература

1. Ожегов С.И.Словарь русского языка:Ок.57 тыс. слов / Под ред.чл.-корр.АНСССР Н.Ю. Шведовой.-17-е изд., стереотип.-М.: Рус.яз.1985.-797с.
2. Г.Ф.Дюбко, Д.В.Преснякова. Модель поверхностного смысла естественного языка на базе семантических функций. //Бионика интеллекта: научн.-техн.журнал 2007. № 1(66). С. 103-106.