УДК 519.7

Н.А. Валенда

Харьковский национальный университет радиоэлектроники, Харьков

МЕТОД АНАЛИЗА МНОГОЗНАЧНЫХ КОНСТРУКЦИЙ ЯЗЫКА

Статья посвящена описанию метода семантического анализа на основе формализованного словаря. Словарь содержит формализованную информацию, которая позволяет выбирать значение слова в зависимости от контекста. Для формализации словаря используются семантические функции. Рассматривается модель установления значения многозначных слов на основе семантических функций и унификации. Рассмотрены алгоритмы формирования семантического представления предложения на основе представлений словосочетаний, входящих в него.

Ключевые слова: *словарь, семантический анализ, контекст, семантическая функция, словосочетание, лингвистический процессор.*

Введение

Обработка постоянно увеличивающихся объемов информации является одной из актуальных задач в области информационных технологий. Значительная часть информации представляется в виде текстов на естественном языке. Задача анализа естественного языка (ЕЯ) является одним из направлений исследований в области искусственного интеллекта (ИИ). Работы по созданию систем анализа ЕЯ начались в 50-х годах 20-го века. Одной из предпосылок, способствовавших возникновению этого направления, была успешная разработка компиляторов для языков программирования, что подталкивало к выводам о возможности применения аналогичных подходов для ЕЯ. Направление ИИ, занимающееся проблемами обработки ЕЯ, получило название Natural Language Processing (NLP).

Основная задача исследований в области обработки ЕЯ – создание эффективных компьютерных моделей ЕЯ. Именно такая постановка задачи отличает NLP от задач традиционной лингвистики и других дисциплин, изучающих ЕЯ, и позволяет отнести ее к области ИИ. Результаты, полученные в работах по данному направлению, могут быть использованы в широком круге систем, взаимодействующих с естественным языком. К ним относятся: системы автоматического перевода и реферирования, поисковые системы, экспертные системы с языковым интерфейсом, системы доступа к базам данных, мультиагентные системы и т.д. Для эффективной работы во всех этих областях требуется моделирование понимания ЕЯ.

Одной из основных задач, которые должны быть решены при автоматическом анализе текстов на ЕЯ, является получение однозначного формального представления. Для решения этой задачи необходимо устранить многозначность, присущую ЕЯ [1, 2]. Многозначность может проявляться на различных уровнях. Морфологическая многозначность может рассматриваться на двух уровнях. Первый уровень – распознавание части речи.

Пример. Слово «рабочий» может являться как существительным, так и прилагательным: новый рабочий, рабочий диск.

Второй уровень – распознавание грамматических категорий, соответствующих одной части речи.

Пример. Неодушевленные существительные мужского рода единственного и множественного числа в именительном и винительном падежах имеют одинаковые формы:

именительный падеж – диск, диски;

винительный падеж – диск, диски.

На уровне синтаксического анализа многозначность проявляется в распознавании роли в предложении.

Пример. Определить подлежащее и дополнение: «программное обеспечение портит вирус».

На уровне семантического анализа многозначность проявляется у слов, имеющих несколько значений.

Пример.

Слово график имеет два значения:

– диаграмма;
– художник.

Получение однозначной интерпретации является важной задачей, решение которой необходимо для реализации систем анализа языка любой сложности.

Семантический словарь

В современных системах анализа языка этап семантического анализа является наиболее трудоемким. Во многих системах, работающих в реальном режиме времени, он представлен в сокращенном варианте. Именно этот этап позволяет повысить качество результирующего представления и получить результат, который моделирует понимание челове-

ком ЕЯ. Сложность семантического анализа является экспоненциальной. Чем длиннее анализируемый отрезок текста, тем больше будет влиять экспоненциальная сложность на скорость работы анализатора ЕЯ. Для сокращения сложности семантического анализа предлагается проводить семантико-синтаксический анализ словосочетаний, входящих в предложение.

Основной составляющей систем обработки текстов на естественном языке является лингвистический процессор. В зависимости от задач, решаемых системой, структура процессора и сложность составляющих элементов могут варьироваться. Основой работы лингвистического процессора являются словари [1].

Использование функционально-семантического подхода основано на применении семантического словаря, содержащего значения слов, которые приписываются им в языке. Такую информацию можно получить из толкового словаря. Для формализации значений слов используется аппарат семантических функций [3]. В электронном словаре каждому слову соответствует отдельная запись, представляющая собой множество суперпозиций семантических функций. Каждая суперпозиция соответствует одному значению слова [4].

Семантический словарь S можно задать в виде объединения множеств – $V_i(x)$ для всех $x \in W(L)$:

$$S = \bigcup_{j=1}^{N} \bigcup_{i=1}^{m} V_i(x_j),$$

где $V_i(x)$ – функция значения слова.

Записи семантического словаря будут иметь следующий вид:

$$S_r = V_i(x) = x \cup f_m(x_1^l,...,x_n^l) \cup \alpha \cup k \cup M(u),$$

где x – заглавное слово;

$f_m(x_1^l,...,x_n^l)$ – суперпозиция семантических функций, соответствующая значению слова x;

α – ссылка на иерархию типов;

k – ссылка на концептуальный граф или пустое множество;

$M(u)$ – множество ссылок на словарь устойчивых словосочетаний или пустое множество.

Модель устранения многозначности лингвистических объектов

Рассматривая лингвистический объект без контекста, невозможно снять омонимию, поскольку нет информации, позволяющей выбрать одну из альтернатив. Чтобы снять омонимию лингвистических объектов, необходимо учитывать контекст. Так, если встречается словосочетание «пушечное ядро», то можно однозначно определить, что слово ядро выступает в первом значении, а в словосочетании «метнуть ядро» – во втором значении. Т.о. метод снятия многозначности должен основываться на анализе контекста. Влияние контекста можно отобразить на примере схемы для двух слов (рис. 1).

Пары вида $b_i^k b_j^l$ задают возможные смысловые сочетания слов a_i и a_j. Только одна из этих пар является истинной для данного высказывания.

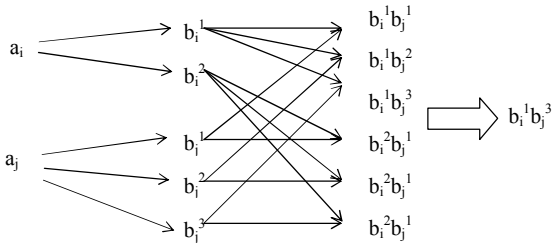

Рис. 1. Влияние контекста на выбор значения

Выбор пары $b_i^1 b_j^3$ означает, что для слова a_i выбрано значение $V_1(a_i)$, а для a_j – значение $V_3(a_j)$. Как видно из рисунка, сложность выбора правильного значения получается экспоненциальной, т.к. осуществляется методом полного перебора. Даже для небольших отрезков текста такой анализ будет весьма емким по времени. Поэтому необходимо проводить его на небольших отрезках текста. Такими отрезками являются словосочетания. Связь между словами должна быть выражена семантическим отношением и формализована с помощью семантической функции. Выбор значения для многозначного слова осуществляться при построении функции словосочетания или заполнении функции предиката. Т.о. количество переборов будет минимальным, а в случае, если только одно слово многозначно – линейным.

Выбор единственного значения лингвистических единиц основывается на алгоритме выбора элемента из множества заполнителей для переменной и алгоритме выбора семантической функции.

Алгоритм выбора элемента из множества заполнителей.

1. Пусть есть переменная – $X_{[P_{morf}, P_{sem}]}$ и множество заполнителей – $\{z_1,...,z_m\}$, будем искать такое z_i, для которого $\sigma = X_{[P_{morf}, P_{sem}]}/z_i$, где z_i может быть словом или языковой конструкцией.

2. Если z_i – слово, производим унификацию переменной и функции значения $V_l(x_j)$. Если z_i – языковая конструкция, производим унификацию переменной и функции значения $V_m(x_1)$. Формируется подстановка $\sigma = X_{[P_{morf}, P_{sem}]}/z_i$.

3. Если для переменной $X_{[P_{morf}, P_{sem}]}$ существует несколько подстановок $\sigma_1,...,\sigma_n$, то произво-

дится унификация семантических функций соответствующих X и z_i, анализируется количество совпавших позиций и выбирается σ_l с наибольшим количеством совпавших позиций.

Алгоритм выбора семантической функции.

1. Если слову соответствует множество значений $x \to \{V_1(x),...,V_n(x)\}$, то для каждого V_i формируется множество переменных семантической функции $\{X^i_{[P_{morf}, P_{sem}]},...,Z^i_{[P_{morf}, P_{sem}]}\}$, где верхний индекс переменной задает принадлежность к функции V_i.

2. Множество заполнителей имеет вид $M = \{V(y_1),...,V(y_k)\}$, где $y_1,...,y_k$ – слова рассматриваемого отрезка текста.

3. Для каждой переменной $Y \in \{X^i_{[P_{morf}, P_{sem}]},...,Z^i_{[P_{morf}, P_{sem}]}\}$ ищем подстановку $\sigma_Y = Y/z$, где $z \in M$, $M = M - z$. Если для всех Y существует подстановка σ_Y, то формируется $M_\sigma = \bigcup_{Y \in M} \sigma_Y$, функция V_i является значением слова x.

4. Процесс повторяется для всех V_i, $i = \overline{1, n}$.

Если не найдено V_i, для которого заполнены все переменные Y, то выбирается V_i с наибольшим числом заполненных переменных.

Метод анализа конструкций естественного языка на основе функционально-семантического подхода

Анализ ЕЯ конструкций предполагает, что в результате некоторой последовательности действий будет построено внутреннее представление на формальном языке, которое будет отображать значение исходной языковой конструкции. Записи семантического словаря представляют собой формализованное описание значения, которое является шаблоном, позволяющим включать в формулу зависимые слова. Каждая функция вырабатывает значение, которое таблично сопоставлено ей в словаре. На основе значения ее можно встраивать в незаполненные позиции другой функции, получая суперпозицию функций. На основании этих свойств построен метод установления зависимостей между лексическими единицами в предложении.

В терминологии грамматик можно сказать, что это анализ снизу вверх – на основании простых единиц строятся более сложные, пока не будет получено предложение. Простейшими составляющими являются слова. Для объединения слов в словосочетания используются семантические функции, соответствующие им в словаре. Основой метода формирования простых словосочетаний является заполнение пустых позиций в суперпозициях семантических функций. На основании простых словосочетаний формируются сложные словосочетания. Основой для формирования словосочетаний является синтаксический анализ (СА), что позволяет не делать полный перебор всех возможных сочетаний слов. Если в предложении есть предикативная часть речи, то полученные словосочетания и слова, которые не вошли в словосочетания, встраиваются в функцию предиката.

Анализ состоит из трех частей:

- формирование простых словосочетаний;
- формирование сложных словосочетаний;
- заполнение функции предиката.

Две первые задачи решаются на этапе синтаксического анализа, третья на этапе семантического анализа.

Формирование простого словосочетания.

1. На основании СА определяется главное и зависимое слово словосочетания – x_1, x_2. Из семантического словаря извлекаем функции, соответствующие главному и зависимому словам: $x_1 \to \{V_1(x_1),...,V_n(x_1)\}$, $x_2 \to \{V_1(x_2),...,V_m(x_2)\}$. Функция словосочетания f строится на основании функции зависимого слова для словосочетаний существительное-прилагательное и на основании функции главного слова для словосочетаний существительное-существительное. Для заполнения переменных будет использоваться значение противоположной функции.

2. Осуществляем выбор семантической функции для функции словосочетания. Если семантические функции содержат переменные, то рассматриваются все возможные сочетания функций словосочетания и множества заполнителей, пока не получим унификацию для аргументов одной из пар. Так формируется функция словосочетания.

3. Если аргументы функций словосочетания не содержат переменных, то формируется стандартная функция на основе синтаксической связи. Ее проверка осуществляется на основании унификации значений входящих в нее слов. Их значения должны пересекаться. Если пересечения нет, то словосочетание не создается.

Формирование сложного словосочетания.

1. В качестве составных частей выступают простые словосочетания и слова. Для анализа используются главные слова простых словосочетаний. На основании СА определяется главное и зависимое слово словосочетания. Выделяется функция словосочетания. Множество заполнителей формируется на основании главных слов словосочетаний или одинарных слов.

2. Осуществляем выбор семантической функции для функции словосочетания. Если семантические функции содержат переменные, то рассматри-

ваются все возможные сочетания функций словосочетания и множества заполнителей, пока не получим унификацию для аргументов одной из пар. Формируется функция словосочетания.

3. Если функция словосочетания не содержит переменных, то формируется стандартная функция на основе синтаксической связи. Ее проверка осуществляется на основании унификации значений входящих в нее слов. Если пересечения нет, то словосочетание не создается.

Формирование языковой конструкции на основе заполнения позиций предиката.

1. Из семантического словаря считываются функции, соответствующие значениям предикативного слова: $x \to \{V_1(x),...,V_n(x)\}$, для каждого $V_i(x)$ создается список переменных $\{X_{1[Pmorf,Psem]}, \ldots, X_{n[Pmorf, Psem]}\}$.

2. Множество заполнителей М формируется из главных слов, сформированных словосочетаний, и одинарных слов.

3. Осуществляется выбор функции предиката на основе алгоритма выбора семантической функции.

4. Если все слова языковой конструкции не могут быть включены в одну формулу, являющуюся суперпозицией функций, то данный отрезок текста семантически не связан. Выполняется процедура выхода.

5. Возвращается суперпозиция семантических функций, построенная на основе функции предиката, свободные позиции которого заполнены словами и словосочетаниями данного отрезка текста.

Выводы

Результаты данной работы могут быть использованы для построения формального представления в системах автоматического перевода, для систем с естественно-языковым интерфейсом, в поисковых системах. Основным преимуществом данного подхода является детальная обработка многозначных слов, что позволяет существенно сократить смысловой недетерминизм для большинства слов и словосочетаний в результирующем формальном представлении.

Разработан метод анализа языковых конструкций на основе заполнения свободных позиций семантических функций. Разработан метод вывода на основе логики предикатов, который отличается от существующих введением унификации для суперпозиций семантических функций. Метод позволяет проводить сопоставление формул, которые отличаются по структуре, но передают близкие значения. Он используется для определения эквивалентности языковых конструкций. Разработана модель выбора значения лексических единиц на основе заполнения свободных позиций и унификации суперпозиций семантических функций.

Список литературы:

1. Хайрова Н.Ф. Автоматизированные информационные библиотечные системы: задачи обработки информации [Текст] / Н.Ф. Хайрова, Н.В. Шаронова. – Х.: Нар. укр. акад., 2003. – 120 с.

2. Хайрова Н.Ф. Машинный перевод [Текст] / Н.Ф. Хайрова, И.В. Замаруева. – Х.: ОКО, 1998. – 80 с.

3. Валенда Н.А. Методы анализа естественного языка на основе функциональной модели семантики [Текст] / Н.А. Валенда, Г.Ф. Дюбко // Бионика интеллекта. – 2005. – № 2 (63). – С. 48-52.

4. Валенда Н.А. Модель функционально семантической обработки текстов на основе унификации [Текст] / Н.А. Валенда // Восточно-Европейский журнал передовых технологий. – 2005. – № 4/2 (16). – С. 95-99.

Поступила в редколлегию 14.12.2010

Рецензент: д-р техн. наук, проф. И.В. Гребенник, Харьковский национальный университет радиоэлектроники, Харьков.

МЕТОД АНАЛІЗУ БАГАТОЗНАЧНИХ КОНСТРУКЦІЙ МОВИ

Н.А. Валенда

Стаття присвячена опису методу семантичного аналізу на основі формалізованого словника. Словник містить формалізовану інформацію, що дозволяє вибирати значення слова залежно від контексту. Для формалізації словника використаються семантичні функції. Розглядається модель установлення значення багатозначних слів на основі семантичних функцій і уніфікації. Розглянуто алгоритми формування семантичного подання речення на основі словосполучень, що входять у нього.

Ключові слова: *словник, семантичний аналіз, контекст, семантична функція, словосполучення, лінгвістичний процесор.*

THE METHOD OF ANALYSIS MULTIVALUED LANGUAGE CONSTRUCTS

N.A. Valenda

The article describes the method of semantic analysis based on a formalized vocabulary. The dictionary contains formalized information that lets you choose the word, depending on the context. Semantic functions are used to formalize a dictionary. We consider a model comparing the values of polysemous words, which is based on semantic functions and unification. Algorithms for the formation of semantic representations of the proposal on the basis of phrases considered in this article.

Keywords*: vocabulary, semantic analysis, context, semantic function, phrase, linguistic processor.*