

Обробка інформації в складних організаційних системах

УДК 519.7

Н.А. Валенда, Л.Д. Самофалов

Харьковский национальный университет радиоэлектроники, Харьков

ФУНКЦИОНАЛЬНО-СЕМАНТИЧЕСКИЙ АНАЛИЗ КОНСТРУКЦИЙ ЕСТЕСТВЕННОГО ЯЗЫКА НА ОСНОВЕ УНИФИКАЦИИ

В работе рассматривается поэтапный анализ предложений естественного языка и получение их представления в виде суперпозиций семантических функций. Данная модель позволяет учитывать многозначность языковых единиц. Частью этого процесса является унификация, которая позволяет сопоставлять значения конструкций естественного языка.

Ключевые слова: унификация, семантический анализ, предикат, семантическая функция, лингвистический процессор.

Введение

Необходимость обработки постоянно увеличивающихся объемов информации на естественном языке определяет актуальность области искусственного интеллекта, занимающейся обработкой текстовой информации.

Обработка информации на естественном языке производится в системах автоматического перевода, поисковых системах, интерфейсах к базам данных, экспертных системах. Результаты работы систем обработки информации характеризуются достаточно низким качеством из-за сложности естественного языка как объекта моделирования. Повышение качества систем обработки информации на естественном языке является одной из актуальных задач в области искусственного интеллекта [1].

Одним из основных подходов к улучшению качества обработки текстовой информации является привлечение семантики и знаний о предметной области. **Целью данной работы** является усовершенствование методов синтаксического и семантического анализа на основе модели семантики в виде семантических функций для повышения качества формального анализа языковых конструкций в системах искусственного интеллекта.

1. Функционально-семантический анализ

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 [2, 3]. В электронном словаре каждому слову соответствует отдельная запись, представляющая собой множество суперпозиций семантических функций. Каждая суперпозиция соответствует одному значению слова.

Анализ конструкций естественного языка (ЕЯ) 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 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 основе 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 основе простых словосочетаний формируются сложные словосочетания. Основой для формирования словосочетаний является синтаксический анализ, что позволяет не делать полный перебор всех возможных сочетаний слов. Если в пред-

ложении 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.

Для формирования простого словосочетания осуществляется следующая последовательность действий.

Синтаксический анализ (СА) определяет главное и зависимое слово словосочетания. Из семантического словаря извлекаем функции, соответствующие главному и зависимому словам. Функция словосочетания f строится на основе функции зависимого слова для словосочетаний существительное–прилагательное и на основе функции главного слова для словосочетаний существительное–существительное.

Для заполнения переменных используется значение противоположной функции.

Если семантические функции содержат переменные, то рассматриваются все возможные сочетания функций словосочетания и множества заполнителей, пока не получим унификацию для аргументов одной из пар.

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.

Для формирования сложного словосочетания осуществляется следующая последовательность действий. В качестве составных частей выступают простые словосочетания и слова. Для анализа используются главные слова простых словосочетаний. На основе СА определяется главное и зависимое слово словосочетания. Выделяется функция словосочетания. Множество заполнителей формируется из главных слов словосочетаний или одинарных слов.

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. Формируется функция словосочетания.

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 семантического представления для предложения формируется на основе функции предиката. Выполняется следующая последовательность действий.

Из семантического словаря считываются функции, соответствующие значениям предикативного слова. Множество заполнителей формируется из главных слов, сформированных словосочетаний, и одинарных слов. Осуществляется выбор функции предиката на основе алгоритма выбора семантической функции.

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

2. Унификация

В исчислении предикатов унификация – это операция, позволяющая привести формулы к общему виду за счет нахождения соответствующих подстановок [4]. Она используется при логическом выводе. Логический вывод предназначен для доказательства следования одной формулы из другой. В данной работе не ставится задача доказательства логического следствия, но должна быть решена другая задача – установление смысловой эквивалентности нескольких формул. Для ее решения формулы надо сопоставить по смысловым значениям элементов, составляющих их. Эта операция подобна унификации в исчислении предикатов, поэтому метод установления смысловой эквивалентности суперпозиций семантических функций будем называть унификацией.

Унификация двух семантических функций означает, что они выражают близкие по смыслу понятия. В работе рассматриваются следующие виды унификации:

- полная унификация семантических функций;
- частичная унификация семантических функций.

Полная унификация – это совпадение структуры и параметров семантических функций. Чтобы функции совпадали, у них должны совпадать функциональные константы и попарно унифицироваться аргументы. Отличия от классического метода заключаются в унификации аргументов. Пары аргу-

ментов выбираются в зависимости от роли, которую они играют в функции. Полная унификация достигается только при совпадении аргументов, которые выступают в одинаковых ролях. При унификации используется понятие подстановки.

Подстановка – отображение множества переменных во множество термов.

Подстановка σ задается множеством пар вида $X/\sigma(X)$, $\sigma(X) \neq X$.

Рассмотрим возможные комбинации аргументов при унификации.

1. Унификация семантической функции f и переменной $X_{[P_{\text{morph}}, P_{\text{sem}}]}$. Если $P_{\text{sem}} \neq \emptyset$, $\exists a_i$, $a_i \subseteq P_{\text{sem}}$, $\exists \alpha$, α – значение главного слова семантической функции f , такие что $a_i = \alpha$ или выполняется $\alpha \leq a_i$, или $a_i \geq \alpha$ для иерархии значений, и $P_{\text{morph}} \neq \emptyset$, $\exists b_i$, $b_i \subseteq P_{\text{morph}}$, $\exists \beta$, β – морфологический признак главного слова семантической функции f , такие что $b_i = \beta$, то данная позиция унифицирована, формируется подстановка $\sigma = X_{[P_{\text{morph}}, P_{\text{sem}}]}/f_i$.

2. Унификация функции значения $V_i(a)$ и переменной $X_{[P_{\text{morph}}, P_{\text{sem}}]}$. Если $P_{\text{sem}} \neq \emptyset$, $\exists a_i$, $a_i \subseteq P_{\text{sem}}$, $\exists \alpha$, α – значение функции $V_i(a)$, такие что $a_i = \alpha$ или выполняется $\alpha \leq a_i$, или $a_i \geq \alpha$ для иерархии значений, и $P_{\text{morph}} \neq \emptyset$, $\exists b_i$, $b_i \subseteq P_{\text{morph}}$, $\exists \beta$, β – морфологический признак a , такие что $b_i = \beta$, то данная позиция унифицирована, формируется подстановка $\sigma = X_{[P_{\text{morph}}, P_{\text{sem}}]}/V_i(a)$.

3. Унификация двух семантических функций $f_i(a^1, \dots, a^k)$, $f_j(b^1, \dots, b^m)$ происходит, если для функциональных констант $i = j$, и аргументы a^1 , b^1 , у которых $l = n$, попарно унифицируются.

4. Унификация двух функций значения $V_i(a)$, $V_j(b)$. Если a и b совпадают, то они унифицируются как константы. В противном случае из семантического словаря берутся суперпозиции семантических функций, которые отвечают индексам функций значения и имеют вид

$$f^{V_i(a)}(x^1, \dots, x^k) = \alpha,$$

$$f^{V_j(b)}(y^1, \dots, y^m) = \beta.$$

Если значения функций совпадают, то аргументы унифицированы. Иначе, если выполняется $\alpha \leq \beta$ или $\beta \geq \alpha$ для иерархии значений, то аргументы унифицированы, в противном случае проводится полная унификация семантических функций, которые входят в суперпозицию.

На практике чаще всего необходимо устанавливать сочетаемость близких по смыслу понятий, а не полную их идентичность. Для этих случаев необходимо ввести понятие частичной унификации. Одно событие может описываться разными семантическими функциями, поэтому даже для идентичных событий семантические функции не совпадут.

Частичная унификация расширяет понятие унификации за счет допуска расхождений в унифицируемых формулах. Расхождения могут быть следующих видов:

- расхождение значений функций;
- расхождение функциональных констант;
- расхождение аргументов функций.

Рассмотрим отдельно каждый случай.

Если функции имеют разные значения, то необходима проверка их соотношения в иерархии значений.

Если одно из значений является потомком другого, то достигается частичная унификация.

Если есть расхождение функциональных констант, то рассматриваются аргументы.

Наиболее важные из отношений, которые задаются семантическими функциями, есть действие, субъект и объект.

Будем считать, если у двух функций перечисленные пары аргументов имеются в наличии и унифицируются, то получаем частичную унификацию.

Расхождение аргументов функций. Если оно есть, то применяются правила, которые были заданы выше для функциональных констант. Отдельно рассматривается случай, который относится только к частичной унификации. Унификация функции значения $V_i(a)$ и семантической функции $f_j(y^1, \dots, y^k)$.

Проводится унификация $V_i(a)$ и $V_j(y^1)$. Если пара функций унифицируема, то исходные функции считаются частично унифицированными. Иначе извлекается семантическая функция соответствующая $V_i(a)$, производится попытка унификации с $f_j(y^1, \dots, y^k)$.

Приведем пример полной унификации языковых конструкций: «Рабочие строят многоэтажный дом», «Строится высотный дом». Для языковых конструкций будут сформированы следующие суперпозиции семантических функций:

$$f_1(V^2(x_1), V^1(x_2), f_4^4(V^4(x_3), V^9(x_4), V^{10}(x_5)))$$

x_1 = строить;

x_2 = рабочие;

x_3 = дом;

x_4 = размер;

x_5 = многоэтажный.

$$f_1(V^2(y_1), V^1(X), f_4^4(V^4(y_2), V^9(y_3), V^{10}(y_4))),$$

y_1 = строить;

y_2 = дом;

y_3 = размер;

y_4 = высотный.

Функциональные константы главных функций совпадают, поэтому унифицируются аргументы.

Первая позиция унифицируется как константы:

$$x_1 = y_1.$$

Вторая позиция унифицируется как переменная и функция значения: $\sigma = X/V(x_2)$.

Третья позиция унифицируется как две семантические функции, для которых совпадает функциональная константа и унифицируются аргументы. Первая и вторая позиции унифицируются как константы: $x_3 = y_2$, $x_4 = y_3$. Для третьей позиции происходит унификация значений на основе иерархии: $x_5 \geq y_4$.

Приведем пример частичной унификации языковых конструкций.

«Строят многоэтажный дом», «Строится небоскреб». Для них будут сформированы следующие суперпозиции семантических функций:

$$f_1(V^2(x_1), V^1(X), f_4^4(V^4(x_2), V^9(x_3), V^{10}(x_4))),$$

x_1 = строить;

x_2 = дом;

x_3 = размер;

x_4 = многоэтажный.

$$f_1(V^2(y_1), V^1(Y), V^4(y_2)),$$

y_1 = строить;

y_2 = небоскреб.

Функциональные символы главных функций совпадают.

Первая позиция унифицируется как константы. Вторая – как две переменные. Третья позиция унифицируется по правилам функции значения и семантической функции. $x_2 \geq y_2$, поэтому достигается частичная унификация.

Выводы

Изложенный подход позволяет объединить мощный аппарат исчисления предикатов и операций унификации с функциональным представлением отношений в конструкциях естественного языка. Полученное представление отражает смысловые отношения в предложении и может быть сопоставлено с другим предложением на основе унификации. Это позволит устанавливать близкие по смыслу предложения, производить детальный анализ предложений для последующего перевода или других видов анализа.

Список литературы

1. Рассел С. Искусственный интеллект: современный подход [Текст] / С. Рассел, П. Норвиг. – М.: «Вильямс», 2006. – 1408 с.

2. Валенда Н.А. Метод формализации семантического словаря на основе функций [Текст] / Н.А. Валенда // Системы обработки информации. – Х.: ХУ ПС, 2013. – Вып. 8 (115). – С. 143-146.

3. Валенда Н.А. Семантико-синтаксический анализ простых конструкций русского языка [Текст] / Н.А. Валенда // Прикладна лінгвістика та лінгвістичні технології: Збірник наукових праць. – Київ, –2009. – С. 104-109.

4. Логический подход к искусственному интеллекту: от классической логики к логическому программированию [Текст]: пер. с англ / А. Тейз, П. Гриббомон, Ж. Луи и др. – М.: Мир, 1990. – 432 с.

Поступила в редколлегию 21.04.2016

Рецензент: д-р физ.-мат. наук, проф. А.В. Грицунов, Харьковский национальный университет радиоэлектроники, Харьков.

ФУНКЦІОНАЛЬНО-СЕМАНТИЧНИЙ АНАЛІЗ КОНСТРУКЦІЙ ПРИРОДНОЇ МОВИ НА ОСНОВІ УНІФІКАЦІЇ

Н.А. Валенда, Л.Д. Самофалов

У роботі розглядається поетапний аналіз пропозицій природної мови і отримання їх уявлення у вигляді суперпозицій семантичних функцій. Дана модель дозволяє враховувати багатозначність мовних одиниць. Частиною цього процесу є уніфікація, яка дозволяє зіставляти значення конструкцій природної мови.

Ключові слова: уніфікація, семантичний аналіз, предикат, семантична функція, лінгвістичний процесор.

FUNCTIONAL-SEMANTIC ANALYSIS OF THE STRUCTURE NATURAL LANGUAGE BASED ON UNIFICATION

N.A. Valenda, L.D. Samofalov

The stage-by-stage analysis of suggestions of human language and receipt of their presentation is in-process examined as superpositions of semantic functions. This model allows to take into account the polisemanticity of linguistic units. Part of this process is standardization which allows to compare the values of human language constructs.

Keywords: standardization, semantic interpretation, predicate, semantic function, linguistic processor.