УДК 519.62

МОДЕЛЬ ПОВЕРХНОСТНОГО СМЫСЛА ЕСТЕСТВЕННОГО ЯЗЫКА НА БАЗЕ СЕМАНТИЧЕСКИХ ФУНКЦИЙ

Г. Ф. Дюбко[1], Д. В. Преснякова[2]

[1] Харьковский национальный университет радиоэлектроники

[2] Харьковский национальный университет радиоэлектроники

В статье предлагается подход к формадизации семантики словосочетаний естественного языка, который декомпозирует семантику на два уровня: поверхностный и глубинный. Рассматривается поверхностный уровень, который представлен семантическими функциями. Приведены определение семантической функции, принцип её использования при доказательстве смысловой эквивалентности словосочетаний, метод конструирования семантических функций из словосочетаний.

ЕСТЕСТВЕННЫЙ ЯЗЫК, СЕМАНТИЧЕСКАЯ ФУНКЦИЯ, ПОВЕРХНОСТНАЯ СЕМАНТИКА, ФОРМАЛИЗАЦИЯ СЕМАНТИКИ.

Введение

Проблема интеллектуального анализа и понимания текстов на естественном языке (ЕЯ) появилась одновременно с созданием компьютеров. Однако до настоящего времени достичь полного успеха в её решении не удалось. Методы для решения частных задач разработаны в большом количестве, но количество так и не перешло в качество. Полное понимание языка на вычислительной (формальной) базе на данный момент остаётся далеко за пределами современных возможностей, и многие фундаментальные проблемы в области автоматической обработки ЕЯ (Natural Language\Processing-NLP) ещё не решены, но развитие современных информационных технологий требует решения частных задач. Так, например, полнотекстовые базы данных, поисковые системы, интерфейсы экспертных систем, машинный перевод требуют анализировать ЕЯ на основе формальных моделей смысла.

Исследователи в области ЕЯ различают общую и прикладную NLP. Задачей общей NLP является разработка моделей обработки языка человеком, а модели при этом должны быть компьютерно эффективными. В основе таких моделей лежит концепция общего понимания текстов и их смысла, как это делается в работах Чарнака, Куллиана, Шенка, Филмора [1, 2, 3, 4].

Прикладная NLP занимается обычно не моделированием, а непосредственной реализацией моделей в некоторой прикладной области. В этом случае не так важно, как реализуемая естественно-языковая конструкция (ЕЯК) будет понята с точки зрения знаний о реальном мире. Здесь важно извлечение информации из ЕЯК о том, чем и как ЭВМ может быть полезной пользователю.

Основной проблемой NLP является языковая неоднозначность: смысловая, падежная, преференциальная, литерация и т. д. Как в общей NLP, так и в прикладной, разрешение неоднозначностей производится с помощью перевода внешнего представления на ЕЯ в некоторую внутреннюю структуру на основе знаний. Для общей NLP такая трансформация требует набора знаний о реальном мире, тогда как для прикладной только об узкой предметной области.

Целью данной работы является построение формальной модели поверхностного смысла на основе функциональной зависимости.

1. Постановка задачи

Смысл ЕЯК является фрагментом знаний, которые содержатся в некоторой форме в соответствующей базе знаний, что требует формальной модели смысла.

Формализация смысла берёт своё начало от работ Н. Хомского. Для отображения смысла используется структура непосредственно составляющих (НС). Преобразование ЕЯК в структуру НС производится посредством порождающей грамматики. Причём в процессе преобразования ЕЯК в НС не используется никакая семантическая информация. В дальнейшем выяснилось, что порождающая грамматика не в состоянии адекватно «языковому миру» выдавать смысл анализируемых конструкций. Одинаковые структуры НС могут приводить к разным смыслам. В то же время в порождающей грамматике не отражено, что разные структуры могут иметь один и тот же смысл. Если дополнить порождающую грамматику моделями, отражающими смысл составляющих ЕЯК, и использовать их в трансформациях, можно достичь большей адекватности модели реалиям «языкового мира».

В данной работе получает дальнейшее развитие модель смысла в виде семантических функций[5], т. е. суперпозиции функций, отражающих смысл словосочетаний языков с фразовой структурой. Суперпозиция функций позволяет композировать смыслы больших ЕЯК из меньших, допуская при этом формальные преобразования конструкций, представляющих смысл. Эта модель должна позволить формальный вывод из ЕЯК, поставляя необходимую информацию о подсмыслах в выводящую систему.

Модель смысла в виде семантических функций должна согласовываться с уже апробированными моделями, предлагая новые возможности (выразительные или вычислительные) и допуская формальное доказательство эквивалентности смыслов языковых конструкций, а также осмысленности анализируемой ЕЯК.

Представление смысла в виде определённой формулы требует вычисления этой формулы посредством программной системы (анализатора ЕЯК), структура которой определяется заданной моделью смысла.

2. Анализатор ЕЯК

Структура предлагаемого анализатора и потоки данных вычислительных блоков представлены на рис. 1.

Рис. 1. Анализатор ЕЯК

Краткая расшифровка функционирования и структур данных представленной схемы приводится ниже.

Блок РИТЛЭ воспринимает ЕЯК и превращает её во внутреннее представление символов, устраняя пробелы и не производя контроля входной информации. Реализовать такой блок можно на уровне конечных автоматов или широко известных алгоритмов поиска подстрок с линейными характеристиками временной сложности алгоритма.

Блок МА превращает поток ЛЭ в последовательность МЕ, т. е. выполняет, по сути, морфологический анализ. Отличием от большинства подходов (особенно в славянских языках) реализации морфологического анализа, где МЕ вычисляется, является декларативный подход, где каноническая форма слова находится в морфологическом словаре. Морфологический словарь в этом случае должен представляться в специфической форме: задаются все словоформы слова и их морфологические признаки. Каноническая словоформа слова связана с такой же словоформой в семантическом словаре. Авторами разработана структура морфологического словаря и эффективный (с вычислительной точки зрения) доступ к канонической форме. Отметим, что элемент МЕ на уровне морфологии может быть неоднозначным. На этом же уровне анализа осуществляется контроль правильности слов во входном потоке.

Блок СА построен на базе использования формальных грамматик и формально организованного семантического словаря. Отличием от традиционного представления атрибутных продукций служит факт включения в продукцию правил применения продукций. Это делает алгоритм анализа более быстродействующим. Обсуждение принципа построения семантического словаря выходит за рамки рассматриваемой здесь работы. Заметим лишь, что словарь построен на основе семантических функций и является лингвистической базой знаний (в отличие от предметной) анализируемого языка. За счёт этой базы знаний возможно существенно сократить нечеткость получаемых знаний. Дальнейшее ограничение нечёткости может быть достигнуто использованием предметной базы знаний.

Блок ФСМ несёт техническую нагрузку, формируя выход анализатора в форме семантических функций из уже сформированных фрагментов.

3. Формальные модели смысла

Большинство работ по формализации семантики языка касаются разработки формальных моделей и языков их представления, которые должны позволять одинаковое формальное каноническое представление для различных ЕЯК, имеющих одинаковый смысл. Формализация семантики ЕЯ уменьшает проблемы, связанные с многозначностью. Само представление в процессе построения канонической формы охватывает много ЕЯ-форм. Однако эквивалентным по смыслу ЕЯ-формам соответствует единственная каноническая форма. Как отмечают исследователи в области языка и математики, про-

цесс преобразования языка в каноническую форму вряд ли можно полностью алгоритмизировать. Как доказано, преобразование в каноническую форму нельзя реализовать даже для моноидов алгебраических групп, которые значительно проще, чем ЕЯ. С точки зрения моделирования языкового поведения человека нет уверенности, что человек хранит свои знания в какой-нибудь канонической форме. В рассматриваемой работе не предлагается канонических форм для представления эквивалентных ЕЯ-конструкций, а доказательство эквивалентности предусмотрено на уровне эвристик.

Анализ литературных источников показал, что формализация семантики сводится преимущественно к построению баз знаний для предметных областей. В этом подходе предложен ряд моделей, которые более или менее сходны с логикой предикатов. Сама логика предикатов является моделью, довольно абстрактно представляющей смысл. К упомянутым моделям относятся семантические сети, концептуальные графы, фреймы, онтологии.

Трансформация ЕЯ-конструкций в смысловые представления на основе баз знаний предметной области встречает определённые трудности вычислительного плана, которые можно уменьшить, декомпозируя трансформацию в два этапа: этап поверхностной семантики и этап глубинной семантики (предметная база знаний). Поверхностный уровень семантики основывается на базе знаний, подобной толковым словарям, используемым человеком.

Как показывают экспериментальные исследования, применение поверхностной семантики позволяет существенно сократить неоднозначность представлений, оставляя возможность установить соответствие между поверхностным и глубинным уровнями.

Предлагаемая в работе модель семантики в виде семантических функций (СФ) относится к поверхностному уровню.

4. Семантические функции

Естественный язык представляется словосочетаниями. Под словосочетанием понимается текст, фраза, сложное предложение, простое предложение, прилагательное плюс существительное, существительное плюс существительное в родительном падеже и т. д. Базовым словосочетанием языка является простое повествовательное предложение. Таким образом, словосочетание — это последовательность слов, составляющих некоторую синтаксическую группу. Каждому словосочетанию ставится в соответствие его смысл. Например, словосочетанию «красный шар» соответствует смысл (речь идёт о поверхностном уровне смысла) «цвет». Это соответствие устанавливается с помощью толкового словаря, где поверхностный смысл описывается в терминах ЕЯ. При этом поверхностный смысл каждого слова задаётся соответствующей словарной статьей этого слова в толковом словаре, а обрабатывает имеющуюся информацию человек-носитель языка.

Чтобы сделать возможной аналогичную человеку обработку языка вычислительными устройствами, необходимо формализовать толковый словарь и представление смысла словосочетаний. Такая формализация может быть достигнута посредством семантических функций. Дадим определение семантической функции.

Обозначим через L(ЕЯ) множество конструкций ЕЯ, т. е. множество словосочетаний, а через L(СФ) — множество конструкций языка, в котором представляется смысл. Заметим, что L(ЕЯ) — неформальный язык, а L(СФ) — формальный. Пусть $\Sigma_т$ терминальный алфавит для представления цепочек в L(ЕЯ), $\Sigma_{рт}$ — часть $\Sigma_т$, используемая только для представления слов L(ЕЯ). Через $\alpha \in \Sigma^*_т$ обозначим языковую конструкцию α, состоящую из символов алфавита $\Sigma_т$ (включая пустое слово Е). Тогда $\Sigma^+_т = \Sigma^*_т - \{E\}$. N — алфавит для представления конструкций языка L(СФ), и $N = \{V, F\_и, (,), ', \} \cup \Sigma_{рт}$.

Пусть $f(\alpha_1, ..., \alpha_n)$ — функция, соответствующая словосочетанию, состоящему из слов $\alpha_1, ..., \alpha_n \in \Sigma^+_т$. Значением этой функции является смысл, который может быть вычислен посредством процедур, приведенных на рис. 1, т. е. $f(\alpha_1, ..., \alpha_n) = \beta$, где $\beta \in N^+$. Языковая конструкция β представляет семантическую функцию

$$F\_и\ (X_1, ... X_m), \qquad (1)$$

где $X_j = V_i(\alpha_k)$ или $X_j = F'\_и$ ($1 \le j \le m$, $1 \le i \le n$, k — номер словарной статьи в семантическом словаре); и — название функции (словосочетания); $F'\_и$ — семантическая функция, вложенная в $F\_и$; α_k — слово из словосочетания.

Следовательно, можно записать:

$$f(\alpha_1, ..., \alpha_n) = F\_и\ (X_1, ... X_m), \qquad (2)$$

где правая часть равенства определена в (1).

Приведём примеры семантических функций в украинском языке. Пусть рассматривается словосочетание «ціна товару в упаковці». Ему соответствует семантическая функция

F_родового відмінку (V_j(ціна), F_прийменник ($V_r(b)$, V_e(товару), V_s(упаковці))).

Для словосочетания «ціна товару в гривнях» семантическая функция имеет вид:

F_родового відмінку (F_прийменник ($V_i(b)$, V_j(ціна), V_k(гривнях) V_e(товару)).

В обоих примерах индексы i, j, k, e, r, s обозначают номер словарной статьи в семантическом словаре.

Выбранные примеры демонстрируют факт, что в славянских языках слова словосочетания могут идти в L(ЕЯ) не подряд, а быть разорванными. Семантическая же функция представляет все слова словосочетания в единой конструкции.

Функция $f(\alpha_1,...,\alpha_n)$ из (2) имеет областью своего определения множество слов, а областью значений – множество строк (термов), являющихся семантической функцией. В свою очередь, семантическая функция может интерпретироваться как функция, аргументами которой являются другие функции, и которая имеет свою область значений. Задавая правила вычисления семантических функций, можно получать их значения. С этой точки зрения равенство (2) можно переписать в виде:

$$f(\alpha_1, ..., \alpha_n) = P^{m+1}F\_и\ (X_1, ... X_m)\ , \qquad (3)$$

где P^{m+1} – $(m+1)$ — местная функция с известными правилами вычисления (вычисляются значения $X_1, ... X_m$, а затем –значение функции $F\_и$).

Областью значений $F\_и$ является множество семантических функций. Возникает вопрос, в чём разница между $F\_и$ и множеством её значений, поскольку значением семантической функции является другая семантическая функция. В ЕЯ смысл некоторой языковой конструкции может отражать конкретику этой конструкции, а может быть обобщённым. Конкретика отражается функцией $F\_и$, а обобщённый смысл значением $F\_и$. Чтобы продемонстрировать это, вернёмся к примеру «красный шар»:

f(красный, шар)=F_прилагательное(V_i(красный), V_j(шар))=F_родительного_падежа(V_k(цвет), V_e(объекта)).

Здесь $F\_$ родительного_падежа является обобщающим смыслом для $F\_$ прилагательное. Заметим, что каждая переменная семантической функции играет определённую роль. Эта роль может фиксироваться либо местом этой переменной, либо ролевым дополнением семантической функции.

Посредством семантических функций можно выразить поверхностную осмысленность ЕЯ-конструкции. Если привлечение семантического словаря не позволяет построить семантическую функцию, то конструкция не обладает смыслом.

Различные семантические функции могут представлять один и тот же смысл. Доказательство этого факта можно формализовать. Рассмотрим это на примере.

Пусть имеются семантические функции

$$F\_\text{глагол}\ (X_1, X_2, X_3), \qquad (2)$$

где X_1 — глагол,обозначающий действие; X_2 — объект, который оказывает действие; X_3 — объект, на который оказывается действие;

$$F\_\text{родительного\_падежа}(Y_1, Y_2), \qquad (3)$$

где Y_1 — существительное; Y_2 -- существительное в родительном падеже или семантическая функция $F\_и$, где существительное в родительном падеже (Y_2) является главным словом.

Функция $F\_и$ есть $F_$творительного падежа, т. е. $F_$творительного падежа (Y_2, Y_3).

Формулы (2) и (3) представляют одинаковый смысл, если слово Y_1(существительное) является отглагольным от глагола X_1, а Y_2 является семантической функцией $F_$творительного падежа с $Y_2=X_2$, $Y_3=X_3$. ЕЯ-конструкции «программа стирает память» и «стирание памяти программой» с семантическими функциями:

$F_$глагол (V(стирать), V(программа), V(память)),
$F_$родительного падежа (V(стирание),
$F_$творительного падежа(V(память),V(программа))),

соответственно, являются эквивалентными по смыслу, что легко проверить по семантическому словарю [6].

С помощью подобных эвристических правил устанавливается эквивалентность для всех возможных словосочетаний.

Выводы

В данном исследовании находит своё отражение подход к формализации смысла словосочетаний естественного языка в виде семантических функций. Семантические функции предлагают формальный аппарат для машинной обработки смыслов, позволяют устанавливать осмысленность и эквивалентность словосочетаний ЕЯ. Точность вычислений регламентируется качеством семантического словаря.

Предложена вычислительная схема для конструирования семантических функций, отражающих смысл конкретных словосочетаний, уточнено определение семантической функции.

Научная новизна работы состоит в том, что получают дальнейшее развитие: понятие семантической функции, метод установления эквивалентности словосочетаний на базе семантических функций, метод построения семантических функций.

Практическая направленность исследований: естественно-языковые интерфейсы, поиск в полнотекстовых базах данных, автоматическое аннотирование и реферирование, машинный перевод.

Дальнейшее направление исследований: выделение всего множества семантических функций в разных естественных языках; определение полного множества операций установления эквивалентности семантических функций; формальные методы представления смысла в семантическом словаре; приложение семантических функций в машинном переводе.

Список литературы: 1. *Charnak E. and Mc Dermot D.* Introduction to Artifical Intellegence. Reading. MA: Addison-Wesly, 1985. — 360 с. 2. *Quillian M.R.* Word concepts. A theory and simulation of some basic semantic capabilities. In Brachman and Levesque. 1967. — 325 с. 3. *Schank R.C. and Rieger C.J.* Inference and the computer understanding of natural language. Artificial Intelligence 5(4): 373–412. 1974. — 512 с. 4. *Filmor C.J.* The case for case. In Bach and Harms,1968. — 420 с. 5. *Дюбко Г.Ф., Валенда Н.А.* Использование формального вывода для семантического анализа конструкций естественного языка // Сборник научных трудов по материалам 6-й международной конференции «Теория и техника передачи, приема и обработки информации». — Х.: ХТУРЭ, 2000. — С. 251–253. 6. *Ожегов С.И.* Словарь русского языка: Ок. 57 тыс. слов / Под ред.чл.-корр. АН СССР Н.Ю. Шведовой. — 17-е изд., стереотип. — М.: Рус. яз., 1985. — 797 с.

Поступила в редколлегию 25.03.07