ИНФОРМАЦИОННАЯ МОДЕЛЬ ОБРАБОТКИ ЕСТЕСТВЕННО-ЯЗЫКОВЫХ ТЕКСТОВ

**Александр Палагин, Виктор Гладун, Николай Петренко,
Виталий Величко, Алексей Севрук, Андрей Михайлюк**

Аннотация: *В статье рассматривается формальная модель обработки естественно-языковых текстов в знаниеориентированных информационных системах. Описаны компоненты, реализующие функции предложенной формальной модели.*

Ключевые слова: *обработка естественно-языковых текстов.*

Архитектура современных знаниеориентированных информационных систем (ЗОС) с естественно-языковым представлением и обработкой знаний включает онтологическую составляющую эксплицитно, которую в общем виде можно интерпретировать как концептуальную базу знаний. Такая база знаний представляется в виде ориентированного графа, вершинами которого являются фреймы, описывающие концепты, а дугами – множество концептуальных отношений, связывающих между собой концепты. Другой важной особенностью указанной архитектуры является разделение и отдельная обработка семантики первой и второй ступени [1], что в общем случае означает разделение внутриязыкового и внеязыкового процессинга [2] и переход к формально-логическому представлению исходного текста.

Указанные особенности архитектуры современных ЗОС трансформируют традиционную модель обработки естественно-языковых текстов (ЕЯТ) в формальную модель следующего вида

$F = \langle T, W, SS^1, O, S^2, I \rangle$, где

T – множество обрабатываемых ЕЯТ;

W – множество словоформ, входящих в T ;

SS^1 – множество синтактико-семантических структур первой ступени, описывающих T ;

O – множество онтологических структур, отображающих множества W и SS^1 в S^2;

S^2 – множество семантических структур второй ступени, описывающих множество сценариев T ;

I – множество информационно-кодовых представлений S^2 .

Опишем объекты формальной модели.

Множество T представляет совокупность естественно-языковых текстов, характеризующихся стилями делового и научно-технического характера.

Цепочка $W \to SS^1$ в классическом понимании представляет грамматический анализ ЕЯТ. В отличие от традиционных линейного и сильнокодированного методов анализа мы используем смешанный метод анализа. Суть его состоит в том, что в лексикографической базе данных полное множество W представлено в таблицах двух типов: таблицами лексем с соответствующими морфологическими, синтаксическими и семантическими характеристиками и таблицами флексий для всех полнозначных, изменяющихся частей речи. При этом алгоритмы формирования парадигмы лексем просты; в таблицах лексики указаны основы лексем и соответствующие коды для выбора записей из таблиц флексий. Нефлексийные изменения учитываются соответствующими алгоритмами.

Описанная структура грамматического анализа однозначно соответствует эффективному отображению функциональных операторов на аппаратный уровень реализации, в частности в базисе ПЛИС (программируемые логические интегральные схемы).

Множество O онтологических структур в идеале представляет языково-онтологическую картину мира, описанную в [1, 3].

Множество SS^1 формируется и интерпретируются итерационно подсистемой синтактико-семантического анализа типа “Конспект” [4]. Основной операцией синтактико-семантического анализа является распознавание синтаксических и семантических отношений, связывающих слова текста. Распознавание связей между знаменательными словами осуществляется путем анализа флексий и предлогов на основе лексических моделей без использования в явном виде правил традиционной грамматики. Для каждого предложения исходного текста строится дерево разбора. Разрешение семантической неоднозначности осуществляется путем обращения к множеству онтологических структур O. На основе построенных деревьев разбора фраз строится категориальная сеть, представляющая собой семантическое пространство S^2 текста. В качестве компьютерного представления такого пространства текста удобно использовать растущую семантическую сеть множества информационно-кодовых представлений I, организованную на основе пирамидальной сети, рецепторы которой соответствуют именам объектов, классов объектов, свойств, состояний, действий, отношений, семантических падежей, модификаторов [5].

Цепочка преобразования информации $T \to W \to SS^1$ и $O \to S^2 \to I$, по сути, представляют (соответственно) базовые процедуры анализа и понимания ЕЯТ, средствами интерпретации которых являются грамматический и семантический процессоры.

В приложениях поиска и обработки большого объема текстовых документов целесообразно использовать знаниеориентированную поисковую систему [6], обеспечивающую начальный и конечный этапы обработки документов - поиска в Интернет и сохранения документов в базе данных в виде их конспектов, сгенерированных подсистемой “Конспект”.

Описанная модель обработки естественно-языковых текстов в знаниеориентированной информационной системе, включающей подсистему “Конспект” как компоненту, представляет перспективное направление развития онтолого-управляемых информационных систем, активно использующих онтологию лексики естественного языка.

Литература

Палагін О.В., Петренко М.Г. Модель категоріального рівня мовно-онтологічної картини світу //Математичні машини і системи. – 2006. - №3. - С.91-104.

Рубашкин В.Ш. Представление и анализ смысла в интеллектуальных информационных системах. – М.: Наука, 1989. – 191с.

Палагин А.В. Организация и функции "языковой" картины мира в смысловой интерпретации ЕЯ - сообщений //Information Theories and Application. – 2000. – Vol. 7, №4. С.155-163.

Гладун В.П., Величко В.Ю. Конспектирование естественно-языковых текстов. Proceedings of the XI-th International Conference “Knowledge-Dialogue-Solution”(KDS’2005).- Varna, Bulgaria.-2005.- pp.344-347 vol.2.

Гладун В.П. Планирование решений. - Киев: Наукова думка, 1987. -168с.

Севрук О.О., Петренко М.Г. Знання-орієнтована пошукова система на основі мовно-онтологічної картини світу //Тези доповідей XIII міжнародної конференції з автоматичного управління “Автоматика-2006”. – Вінниця. – 2006. - 25-28 вересня. – С.413.

Информация об авторах

Палагин Александр Васильевич - Ин-т кибернетики им. В.М. Глушкова НАН Украины, Киев-187 ГСП, 03680, просп. акад. Глушкова, 40,e-mail: palagin_a@ukr.net

Гладун Виктор Поликарпович - Ин-т кибернетики им. В.М. Глушкова НАН Украины, Киев-187 ГСП, 03680, просп. акад. Глушкова, 40,e-mail: glad@aduis.kiev.ua

Петренко Николай Григорьевич - Ин-т кибернетики им. В.М. Глушкова НАН Украины, Киев-187 ГСП, 03680, просп. акад. Глушкова, 40,e-mail: petrng@ukr.net

Величко Виталий Юрьевич - Ин-т кибернетики им. В.М. Глушкова НАН Украины, Киев-187 ГСП, 03680, просп. акад. Глушкова, 40,e-mail: glad@aduis.kiev.ua

Севрук Алексей Олегович, Михайлюк Андрей Васильевич - Ин-т кибернетики им. В.М. Глушкова НАН Украины, Киев-187 ГСП, 03680, просп. акад. Глушкова, 40,e-mail: petrng@ukr.net