ОБРАБОТКА ПРЕДЛОЖЕНИЙ ЕСТЕСТВЕННОГО ЯЗЫКА С ИСПОЛЬЗОВАНИЕМ СЛОВАРЕЙ И ЧАСТОТЫ ПОЯВЛЕНИЯ СЛОВ

Александр Палагин, Сергей Крывый, Дмитрий Бибиков

***Аннотация**: Описывается один из подходов к анализу естественно-языкового текста, который использует толковый словарь естественного языка, локальный словарь анализируемого текста и частотные характеристики слов в этом тексте.*

***Ключевые слова:** представление текста, обработка текста, формальная логическая система.*

***ACM Classification Keywords**: I.2.4 Knowledge Representation Formalisms and Methods - Representation languages, I.2.7 Natural Language Processing - Language models*

***Conference**: The paper is selected from XV[th] International Conference "Knowledge-Dialogue-Solution" KDS 2009, Varna, Bulgaria, June-July 2009*

Введение

Проблемы, связанные с анализом естественно-языковых объектов (ЕЯО) традиционно относят к области искусственного интеллекта. Однако, объективные трудности, возникающие на пути анализа ЕЯО, не позволяют удовлетворительно решать проблему автоматизации такого анализа. Эти трудности связаны с тем, что проблема анализа ЕЯО относится к проблемам, которые плохо поддаются формализации. По проблеме анализа ЕЯО существует огромная литература, в которой описываются различные методы и подходы к решению частных случаев данной проблемы [см. 1-5]. В данной работе решается задача семантического анализа предложений естественного языка с целью извлечения знания, имеющегося в анализируемом тексте. При этом семантический анализ ограничивается одним предложением, и не рассматривается вопрос связей между предложениями, хотя некоторые предположения на этот счет делаются.

1. Общая постановка задачи анализа ЕЯО

Описываемая ниже общая постановка задачи анализа ЕЯО была сформулирована в работе [6] и здесь она конкретизируется для решения анализа предложений естественного языка.

Пусть $T = t_1 t_2 \ldots t_n$ естественно-языковый текст в алфавите X, т.е. $T \in L(X)$, где $L(X)$ - язык над алфавитом X, а $t_i \in T$ - предложения, $i = 1,2\ldots,n$.

Каждое предложение $t_i \in T$, в свою очередь, имеет структуру $t_i = t_{i_1} t_{i_2} \ldots t_{i_m}$, где t_{i_j} содержательно означают грамматические единицы, из которых построено предложение t_i. Если $t_{i_j} \in t_i$, то $C_L(t_{i_j}) = t_{i_1} \ldots t_{i_{j-1}}$ и $C_R(t_{i_j}) = t_{i_{1+1}} \ldots t_{i_m}$ будем называть левым и правым контекстом слова t_{i_j} соответственно в предложении t_i.

С текстом T свяжем такие объекты:

- S - словарь языка $L(X)$, где содержатся слова t_{i_j} со своими определениями;
- $\gamma \subseteq T \times S$ - отношение, определяющее возможные значения и типы слова в словаре S;

- $A = (D, \Pi)$ - предметная модель, на которой интерпретируется текст T;
- $\phi \subseteq T \times A$ - отношение интерпретации текста T на модели $A = (D, \Pi)$.

Сигнатура предикатов $\Pi = \{\pi_{k_1}, \ldots, \pi_{k_r}\}$ содержит атомарные предикаты, из которых можно строить сложные формулы. Сейчас мы не будем фиксировать эту сигнатуру, поскольку она зависит от предметной модели. В связи с тем, что модель не уточняется, то и её сигнатуру уточнить нельзя. Заметим только то, что каждый атомарный предикат имеет тип (т.е. это будет некоторая типизированная сигнатура).

Определим теперь правила вычисления отношений γ и ϕ.

Отношение γ имеет достаточно простой способ вычисления:

$$\gamma(t_{i_j}) = \{(a_1, \tau_1), (a_2, \tau_2), \ldots, (a_s, \tau_s)\},$$

где a_i возможные значения слова t_{i_j}, а τ_i - его возможные типы. Может случиться, что $\gamma(t_{i_j}) = \varnothing$. В этом случае значение этого слова считается неопределенным (и это требует пополнения словаря S).

Отношение ϕ определяется несколько сложнее. Если модель $A = (D, \Pi)$ определена, то

$$\phi(T) = \phi(t_1) \ldots \phi(t_n), \text{ где}$$

$$\phi(t_i) = \{\phi(\gamma(t_{i_1})\gamma(C_R(t_{i_1}))), \phi(\gamma(C_R(t_{i_2}))\gamma(t_{i_2})\gamma(C_R(t_{i_2}))), \ldots, \phi(\gamma(C_L(t_{i_n}))\gamma(t_{i_n}))\},$$

при этом

$$\phi(\gamma(t_{i_j})) = \gamma(\phi(t_{i_j}));$$

$$\phi(\gamma(C_L(t_{i_j}))) = C_L(\gamma(\phi(t_{i_j}))));$$

$$\phi(\gamma(\pi_r^k(p_1, \ldots, p_k))) = \gamma(\phi(\pi_r^k))(\phi(\gamma(p_1), \ldots, \gamma(p_2))),$$

где $\gamma(\phi(\pi_r^k))$ - имя предиката, тип которого согласован с аргументами $\gamma(p_1), \ldots, \gamma(p_k)$.

Из этой формальной постановки проблемы анализа ЕЯО вытекает, что основные задачи сводятся к таким:

- построить предметную модель A; эта задача является основной и наиболее трудной в связи с тем, что предметная модель является по существу базой знаний (построение такой базы состоит в том, чтобы определиться с объектами, которые извлекаются из текста, с формальным логическим языком, правилами вывода, аксиоматикой и пр.);
- показать вычислимость отношений γ и ϕ на предметной модели A и построить алгоритмы вычисления отношений γ и ϕ;
- при вычислении отношений γ и ϕ контролировать соответствие типов аргументов и предикатов;
- определить взаимодействие алгоритмов вычисления γ и ϕ с системами синтаксического и семантического анализа текста.

Второстепенными, но тоже важными, являются задачи связанные:

- с определением структуры данных для словарей;
- с определением информации, которая должна содержаться в словарях;
- с определением режима взаимодействия с пользователем (автоматический, полуавтоматический, диалоговый);
- язык пользовательского интерфейса и алгоритмы логического вывода.

2. Конкретизация общей задачи и схема ее решения

Нормализация ЕЯО. Очевидно, что в такой постановке проблема анализа ЕЯО носит очень общий характер, поэтому задачу необходимо конкретизировать. Рассмотрим два подхода к такой конкретизации. Большинство систем как информационного поиска, так и обработки текстовой информации содержат в качестве основного компонента систему анализа, служащую для выявления «содержания» или «значения» заданной единицы информации. В обычных системах такого рода анализ может выполнять человек. При этом он использует заранее разработанные таблицы или шаблоны для определения того, какой идентификатор содержания больше подходит по смыслу для заданной единицы информации. Известны также системы так называемого автоматического индексирования, в которых идентификаторы содержания приписываются автоматически, исходя из структуры текста документа и запроса.

В связи с тем, что естественный язык содержит различного рода нерегулярные явления, которые наблюдаются как в синтаксисе, так и в семантике, то система смыслового анализа должна приводить входные тексты к некоторому нормализованному виду, преобразуя различные, возможно неоднозначные, структуры на входе в фиксированные, стандартные идентификаторы содержания. Такого рода процедуры нормализации языка часто используют словари и списки слов, содержащие допустимые идентификаторы содержания, причем для каждого идентификатора приводится соответствующее определение с тем, чтобы регулировать и контролировать его использование. Следует заметить, что до появления понятия «онтология» процедуры анализа ЕЯО редко выходили за рамки анализа одного предложения. Это объясняется тем, что проблема анализа ЕЯО очень сложна и приходится сильно ограничивать свои запросы при попытке автоматизации такого рода анализа, выполняя некоторый упрощенный анализ текста. Рассмотрим такое упрощение, описываемое в данной работе.

Конкретизация задачи анализа. Конкретизации задачи анализа в нашем случае сводится к следующему.

Словарь S, о котором говорилось выше, является толковым словарем языка $L(X)$ (это может быть словарь русского, украинского, английского или какого-либо другого естественного языка).

Текст T состоит из предложений языка $L(X)$ и представляет текст, который не содержит никаких символов, кроме символов алфавита X (т.е. T не содержит формул, графиков, рисунков и т.п.).

Отношение γ состоит из суперпозиции двух отношений $\gamma_1 * \gamma_2$, выполняемых последовательно. Содержательно отношение γ_1 означает распознавание принадлежности слова к данному языку и проверку правильности написания слова $t_{i_j} \in t_i$, где $t_i \in T$ в соответствии с написанием его в толковом словаре, т.е.

$$\gamma_1(t_{i_j}) = \begin{cases} 1, & \text{если } t_{i_j} \in S; \\ 0, & \text{если } t_{i_j} \notin S. \end{cases}$$

Если слово $t_{i_j} \in t_i$ распознано в словаре *S*, то оно заносится в словарь *T'* правильных слов, а если это не так, то предусматривается сигнализация о том, что данное слово отсутствует в словаре *S* и принимается решение о добавлении данного слова в словарь или его исправлении (слово может быть искажено, например, вследствие сканирования текста *T*).

Словари *S* и *T'* являются входными данными для отношения γ_2. Содержательный смысл отношения γ_2 сводится к тому, что если $\gamma_1(t_{i_j}) = 1$, то $\gamma_2(t_{i_j})$ определяет его грамматическую единицу языка (имя

собственное, сказуемое, существительное, числительное и т.п.) а также возможные флексии слова $t_{i_j} \in t_i$.

Областью интерпретации текста T является модель $A = (D, \Pi)$, где T - это исходный текст, возможно расширенный некоторой дополнительной информацией, а сигнатура П определяется из текста T в результате использования информации о различных вхождениях слова t_{i_j} в предложения $t_i \in T$. При этом вычисление отношения ϕ ограничивается отдельно взятым предложением $t_i \in T$, определяемым каждым вхождением слова t_{i_j} в текст T. В случае трудности определения предиката $\pi_i \in \Pi$, предусматривается диалоговый режим вычисления $\phi(\pi_i)$ и $\gamma(\phi(\pi_i))$.

Схематично предлагаемая система анализа выглядит следующим образом:

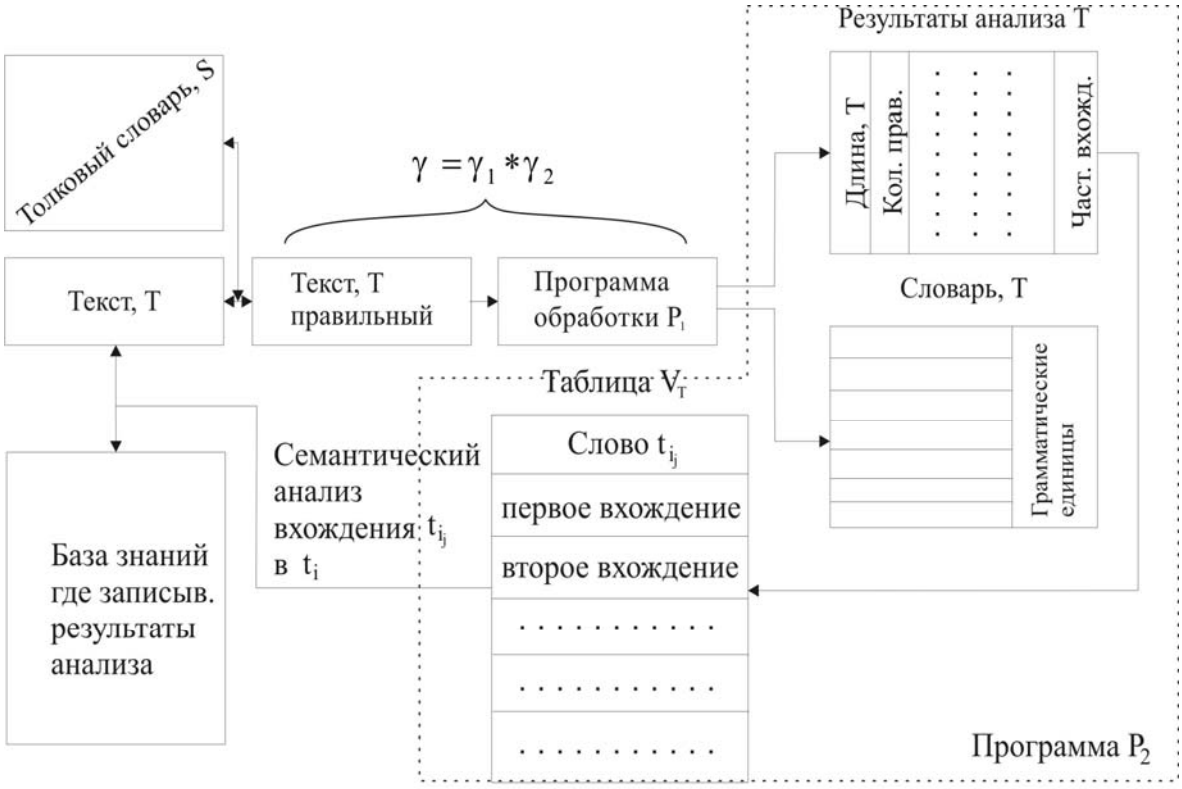
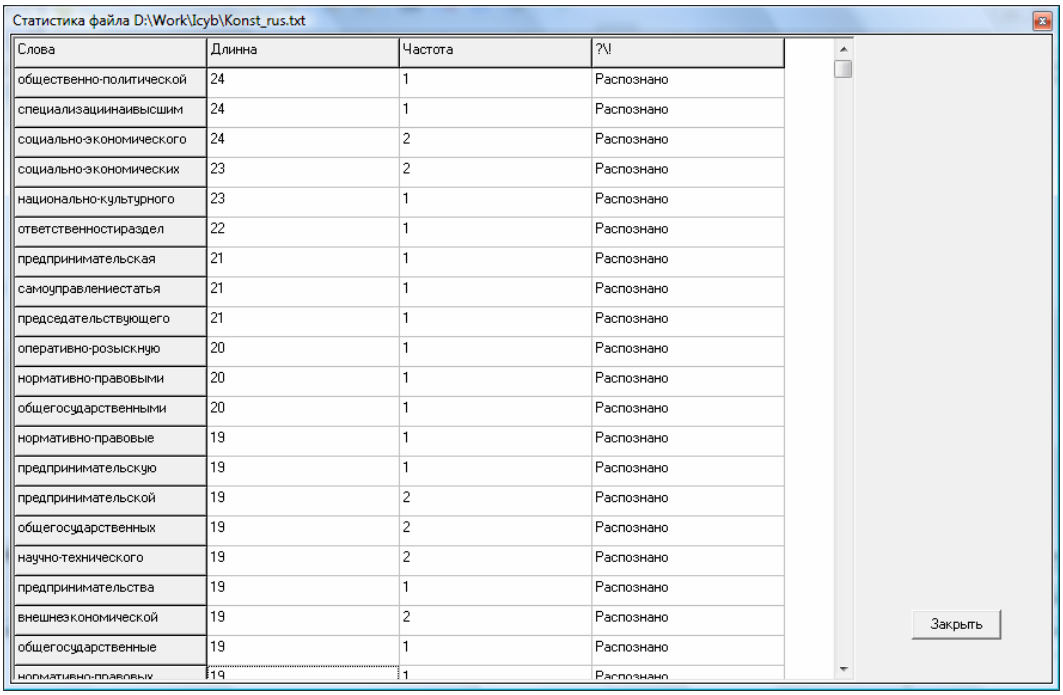


Рис.1. Схема системы анализа предложений естеств. языка

В этой схеме отношение $\gamma = \gamma_1 * \gamma_2$ вычисляет программа P_1. Результатом ее работы является два файла F_1 и F_2, заполненные соответственно числовыми характеристиками слов входного текста T и словами t_{i_j} предложений этого текста. Структура файла F_2 показана ниже на рисунке 2.


Статистика файла D:\Work\Icyb\Konst_rus.txt

Слова	Длинна	Частота	?\!
общественно-политической	24	1	Распознано
специализациинаивысшим	24	1	Распознано
социально-экономического	24	2	Распознано
социально-экономических	23	2	Распознано
национально-культурного	23	1	Распознано
ответственностираздел	22	1	Распознано
предпринимательская	21	1	Распознано
самоуправлениестатья	21	1	Распознано
председательствующего	21	1	Распознано
оперативно-розыскную	20	1	Распознано
нормативно-правовыми	20	1	Распознано
общегосударственными	20	1	Распознано
нормативно-правовые	19	1	Распознано
предпринимательскую	19	1	Распознано
предпринимательской	19	2	Распознано
общегосударственных	19	2	Распознано
научно-технического	19	2	Распознано
предпринимательства	19	1	Распознано
внешнеэкономической	19	2	Распознано
общегосударственные	19	1	Распознано
нормативно-правовых	19	1	Распознано

Закрыть

Рис. 2. Структура файла F_2

Файлы F_1 и F_2, сформированные программой P_1, служат входными данными для работы программы P_2, которая вычисляет отношение ϕ. При этом, работа программы P_2 сводится к построению таблицы V_T для слов $t_{i_j} \in t_i$, $t_i \in T$. Затем, по этой таблице и предложениям текста T определяется семантический смысл рассматриваемого предложения. Предложение $t_i \in T$ определяется на основе номера вхождения слова t_{i_j} в текст T с помощью таблицы V_T, вид которой приведен ниже на рисунке 3..

Слово	№ вхождения	№ предложения, t_i
Палагин	1	1
	2	2
	3	3
	4	5
Кривой	1	1
	2	4
Петренко	1	1
	2	2
Яковлев	1	3
Опанасенко	1	5
Кургаев	1	6

Рис. 3. Структура таблицы V_T

3. Пример работы системы

Рассмотрим в качестве примера текста T текст некоторой библиографической информации и анализ этого текста с помощью описанной системы (например, тексты статей, присылаемых на конференцию, и их библиография). Такого рода текст имеет четко нормализованную структуру: нулевая позиция – универсальный определитель (УДК), первая часть текста – список авторов, вторая часть – название статьи, третья часть – издательство, четвертая – год издания.

Тогда построение модели $A = (D, \Pi)$ сводится к наполнению данными множеств УДК, АВТОР, СТАТЬЯ или КНИГА, ИЗД-ВО, ГОД, РЕЦ, где множество РЕЦ означает множество рецензентов, которое предположительно имеется и построено пользователем системы. Это наполнение в данном случае выполняется автоматически, с использованием имеющейся таблицы V_T. Более того, использование информации о вхождении того или иного слова в текст T позволяет автоматически построить декартовое произведение

$$D = УДК \times АВТОР \times (СТАТЬЯ \cup КНИГА) \times ИЗД\text{-}ВО \times ГОД \times РЕЦ.$$

Построив это множество, можно строить сигнатуру предикатов, хотя в данном случае в этом нет необходимости. Это объясняется тем, что из множества *D* средствами реляционной базы знаний можно построить отношения, которые нас интересуют. Например, используя селекцию проекции множества *D* по атрибутам АВТОР, СТАТЬЯ, РЕЦ получаем отношение-предикат R_{acp}, содержащее информацию о статьях, авторах и рецензентах статей этих авторов. Приведем конкретные значения текста и результатов его анализа.

Пусть исходный текст представляет список литературы и список РЕЦензентов,

которые рецензировали эти работы:

1. Палагин А.В., Кривой С.Л., Петренко Н.Г. статья1. - изд1. - 2009. - С.1-10. - (рец. Р1).

2. Палагин А.В., Петренко Н.Г. статья2. - изд2. - 2007. - С. 21-30. - (рец. Р2).

3. Палагин А.В., Яковлев Ю.С. книга1. - изд3. - 2002. - 500 с. - (рец. Р3).

4. Кривой С.Л. статья3. - изд3. - 2007. - С. 5-17. - (рец. Р4).

5. Палагин А.В., Опанасенко В.Н. книга2. - изд4. - 2004. – 300 с. - (рец. Р2).

6. Кургаев А.Ф. книга 3. - изд5. - 2007. - 450 с. - (рец. Р3).

Анализ этого текста дает такие множества предметных констант и их семантические значения:

АВТОР = {Палагин А.В., Кривой С.Л., Петренко Н.Г., Яковлев Ю.С.,Опанасенко В.Н., Кургаев А.Ф.},

СТАТЬЯ $\cup$ КНИГИ = {статья1, статья2, статья3, книга1, книга2, книга3},

ИЗД = {изд1, изд2, изд3, изд4, изд5},

РЕЦ = {Р1, Р2, Р3, Р4}.

Применяя оператор декартового произведения отношений реляционной алгебры, получаем множество *D*. Из этого множества тем же способом с помощью операторов проекции и селекции получаем отношения:

R_a = { *Палагин А.В., Кривой С.Л., Петренко Н.Г., Яковлев Ю.С., Опанасенко В.Н., Кургаев А.Ф.*},

элементы которого упорядочены так, как они идут в этом отношении. Возможная интерпретация констант в словаре *S* может выглядеть так:

γ *(автор1) = (Палагин А.В., имя-собственное, дтн, профессор, академик,...)*

Аналогично по остальным авторам.

Далее из множества *D* определяются отношения-предикаты: отношение "автор-статья"

$$R_{ac} \subseteq \text{АВТОР} \times (\text{СТАТЬИ} \cup \text{КНИГИ}),$$

семантическое значение которого имеет вид:

R_{ac} ={*Палагин А.В., статья1), (Палагин А.В., статья2),(Палагин А.В., книга1),*
(Палагин А.В., книга2), (Кривой С.Л., статья1), (Кривой С.Л., статья3),
(Петренко Н.Г., статья1), (Петренко Н.Г., статья2), (Яковлев Ю.С., книга1),
(Опанасенко В.Н., книга2), (Кургаев А.Ф., книга3)};

отношение "автор-рецензент-статья"

$$R_{acp} \subseteq \text{АВТОР} \times \text{РЕЦ} \times \text{СТАТЬИ} \cup \text{АВТОР} \times \text{РЕЦ} \times \text{КНИГИ},$$

семантическое значение которого имеет вид:

R_{acp} ={ *(Палагин А.В., Р2, статья1), (Палагин А.В., Р1, статья2),*
(Палагин А.В., Р2, книга1), (Палагин А.В., Р3, книга2),
(Палагин А.В., Р3, книга1), (Палагин А.В., Р1, книга2),
(Палагин А.В., Р4, книга1), (Палагин А.В., Р2, книга2),
(Кривой С.Л., Р4, статья1), (Кривой С.Л., Р4, статья3),
(Петренко Н.Г., Р2, статья1), (Петренко Н.Г., Р3, статья2),
(Яковлев Ю.С., Р2, книга1), (Опанасенко В.Н., Р3, книга2),
(Яковлев Ю.С., Р3, книга1), (Опанасенко В.Н., Р4, книга2),
(Яковлев Ю.С., Р4, книга1), (Опанасенко В.Н., Р1, книга2),
(Кургаев А.Ф., Р1, книга3), (Кургаев А.Ф., Р2, книга3),
(Кургаев А.Ф., Р4, книга3)}.

Первое отношение бинарное, а второе - тернарное. Аналогичным способом можно получить и другие интересующие пользователя отношения-предикаты, т.е. полностью построить сигнатуру предикатов *П,* исходя из области *D*.

Заключение

Данная работа посвящена проектированию автоматизированных систем извлечения и обработки знаний, а также онтологий предметных областей. Она является продолжением работ [6-8], где описывалась технические возможности реализации такого типа систем.

Какие задачи можно решать с помощью предлагаемой системы? Поскольку данная система является только первым этапом на пути создания автоматизированных систем извлечения и обработки знаний, то эта система извлекает информацию из ЕЯО, необходимую для построения онтологий предметной области, к которой относится анализируемый текст. А именно,

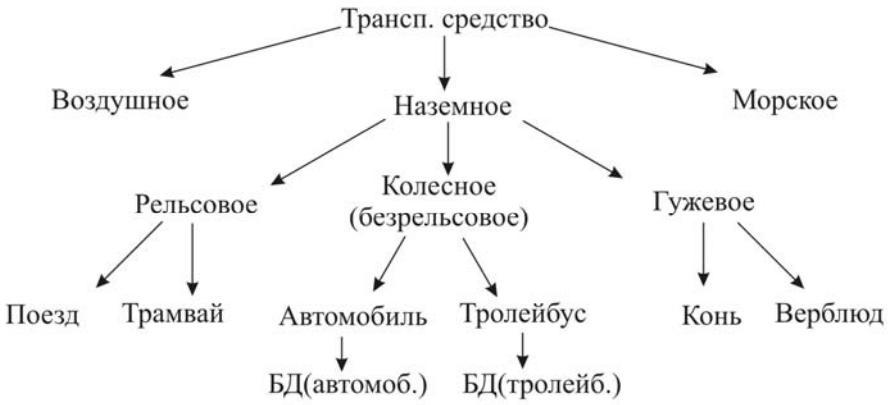
а) распознается смысл данного конкретного слова в зависимости от того, в какое предложение оно входит;

б) строится модель $A = (D, \Pi)$, элементы которой разбиваются по типам (концепты, предикаты, константы и т. п.). После такого разбиения эта модель является основой для построения онтологии данной предметной области или пополнения уже существующих онтологий;

в) позволяет ввести время, в котором происходят события в данном тексте.

Следующим шагом на пути анализа ЕЯО является использование определений с целью построения онтологии предметной области или пополнения уже существующей. Эта идея была предложена в работе [9], которая базировалась на использовании теории конечных автоматов. Пополнение имеющейся онтологии связывалось с операцией итерации, а соединение нескольких онтологий в одну (интеграция онтологий) – с операциями объединения, пересечения и конкатенации автоматов, представляющих данные онтологии. Для пояснения этого подхода рассмотрим пример.

Пусть строится онтология **«Транспортные средства»**. Известно, что транспортные средства делятся на наземные, морские и воздушные. В свою очередь наземные транспортные средства делятся на рельсовые, колёсные (безрельсовые) и гужевые. К рельсовым транспортным средствам относится поезд, трамвай, к безрельсовым – автомобиль, автобус, троллейбус, к гужевым – верблюд, лошадь. Эта классификация дает возможность построить такую онтологию:



Пусть в тексте встречаются определения:

Троллейбус – безрельсовое транспортное средство, приводимое в движение электродвигателем.

Автобус – колесное транспортное средство, *приводимое* ***в движение двигателем внутреннего сгорания.***

Тогда система анализа строит такую конструкцию:

$$Тролейбус(x) \Leftrightarrow Трансп.средство(x) \wedge \neg(Рельсовое) \wedge \exists y(Прив-в-Движ(x,y) \wedge Электродв(y))$$

По этой конструкции система находит в имеющейся онтологии концепты «Трансп. средство», «(безрельсовое) колесное» и «троллейбус». Отсутствие в данной онтологии концепта «электродвигатель» приводит к тому, что система ищет онтологию «Двигатели» или «Электродвигатели». Если находит, то строится предикат $\exists y(Прив-в-Движ(x,y))$, где x и y конкретизированные концепты-переменные. В противном случае система сигнализирует о неполноте существующей онтологии, которую необходимо пополнить понятием «Двигатели». После пополнения имеющейся онтологии, анализ второго определения уже не приводит к ситуации неполноты имеющейся онтологии.

Заметим, что система анализируя определение, естественным образом находит иерархию понятий. Например, в определении троллейбус понятие “Трансп.средство” является более общим, чем понятие «безрельсовое». Отсюда следует, что нормализация текста должна сводится к тому, что в определениях отношение следования необходимо задавать от «более общего» к «более конкретному».

Благодарности

Работа выполнена при финансовой поддержке Министерства образования и науки Украины в рамках совместного Украинско-Болгарского проекта № **145 / 23.02.2009** «Разработка распределенных виртуальных лабораторий на основе прогрессивных методов доступа для поддержки проектирования сенсорных систем» и Болгарского национального научного фонда в рамках совместного Болгарско-Украинского проекта **D 002-331 / 19.12.2008** с тем же названием.

Литература

1. Кулик Б.А. Логика естественных рассуждений. С.-П.: Невский диалект.-2001.-127 с.
2. Рубашкин В.Ш. Представление и анализ смысла в интеллектуальных информационных системах.-М.:Наука.-1989.-191с.
3. Тейз А., Грибомон П., Юлен Г. И др. Логический подход к искусственному интеллекту: От модальной логики к логике баз данных.-М.: Мир.-1998.-492с.
4. Широков В.А. Феноменологія лексикографічних систем. К.:Наукова думка.-2004.-327с.
5. Апресян Ю.Д. и др. Лингвистический процессор для сложных информационных систем.-М.: Наука.-1992.-287с.
6. Палагин А.В., Крывый С.Л., Петренко Н.Г. Знаниеориентированые информационные системы обработки знаний, содержащихся в естественно-языковых объектах: методологические основы и архитектурная организация. – УсиМ.-2009. (в печати)
7. Палагин А.В. Архітектура онтологоуправляемых компьютерных систем. - Киберн. и сист. анализ.- 2006.-№2.-С.111-124.
8. Палагін О.В., Петренко М.Г. Архітектурно-онтологічні принципи побудови інтелектуальних систем. - Математичні машини і системи.-2006.-№4.- С.15-20.
9. Крывый С.Л., Ходзинский А.Н. Автоматное представление онтологий и операции на онтологиях. –International Book Series. – N 1. – Algorithmic and Mathematical Foundations of the Artificial Inetligence. – ITHEA: Sofia. -2008. – PP. 173-179.

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*

Крывый Сергей Лукъянович – *Киевский национальный университет им. Т.Г. Шевченко, Киев, ГСП, 01601, ул. Владимирская, 64, e-mail: krivoi@i.com.ua*

Бибиков Дмитрий Сергеевич *– Ин-т кибернетики им. В.М. Глушкова НАН Украины, Киев-187 ГСП, 03680, просп. акад. Глушкова, 40, e-mail:* bb_coff@mail.ru