

УДК 164.053

**ОБНАРУЖЕНИЕ СМЫСЛА ТЕКСТОВ ЕСТЕСТВЕННОГО ЯЗЫКА
ПРИ ПОМОЩИ КОМПЬЮТЕРНЫХ ВЫЧИСЛЕНИЙ**

Л.А. Святогор, В.П. Гладун

Институт кибернетики им. В.М. Глушкова НАН Украины

e-mail: aduis@rambler.ru

Введение

Человек, существуя в динамичной социальной среде и находясь при этом в сознании, постоянно размышляет. Значительные объёмы информации он получает из литературных источников. В процессе чтения текста или дискурса потребитель стремится, следуя авторскому замыслу, представить ситуацию в контексте других событий, постичь её локальный смысл и построить общий смысл. Только уяснив этот смысл для себя лично, потребитель может его запомнить, пересказать, прогнозировать развитие событий или даже развить саму идею. Такие когнитивные процессы познания связывают с «пониманием» [1].

Цель данной работы состоит в том, чтобы выяснить, какой должна быть интеллектуальная машина, способная при помощи неинтеллектуальных операций (алгоритма) извлекать из текста художественного происхождения или дискурса его смысл. При этом смысл рассматривается как некоторый фрагмент знания, а его формальное определение опирается на структурное представление знаний.

Система репрезентации знаний

Среди семантических сетей особое место занимает *граф концептуальных зависимостей*, предложенный Р. Шенком (1975). Этот граф образует понятийную систему и благодаря этому может быть использован для представления знаний о Море. Концептом называют информацию о предметах и явлениях, которую выражает *имя* [2].

На базе концептуального графа построена **иерархическая трёхуровневая онтология ИО*3**, описанная в [3]. В общем случае онтология – это семантическая сеть с концептами **X**, отношениями **R** и системой интерпретации концептов и отношений – **D**: **O = < X, R, D >**. **Формальными особенностями** онтологии **ИО*3**, которые отличают её от других подобных систем репрезентации знаний, являются следующие: (1) это предметная лексическая онтология, концептами $x \in X$ которой служат имена; (2) имена соответствуют понятиям разной степени обобщения и отображают денотаты внешнего мира; (3) набор связей **R** включает в себя таксономические, атрибутивные и ассоциативные отношения; (4) структура **XX** онтологической сети образует иерархию пирамидального типа: чем выше уровень понятия, тем больше его степень обобщения; (5) систему интерпретации **D(X)** и **D(R)** образуют письменные и электронные источники баз знаний; (6) нижний уровень онтологии может безгранично дотраиваться конкретными предметными областями (ПрО) знаний; (7) онтология взаимодействует с текстом через совпадающие имена **x**.

В общем виде трёхуровневая онтология **ИО*3** представлена на **рисунке 1**. Это – теоретическая модель Знаний. Иерархическая структура модели присуща, по-видимому, организации знаний в памяти человека, который упорядочивает сложную «картину Мира» при помощи обобщений и классификаций.

«Пирамида знаний» формируется инженерами по знаниям и экспертами. При этом ни само множество понятий ПрО, ни объёмы их знаний не могут быть зафиксированы: онтология **ИО*3** динамична во времени и расширяема. Являясь универсальной моделью знаний, данная система в то же время допускает синтез её конечных подсистем, ориентированных на решение конкретных задач семантического анализа, и приспособлена

для компьютерных процедур смысловой обработки текстов естественного языка (ЕЯ).

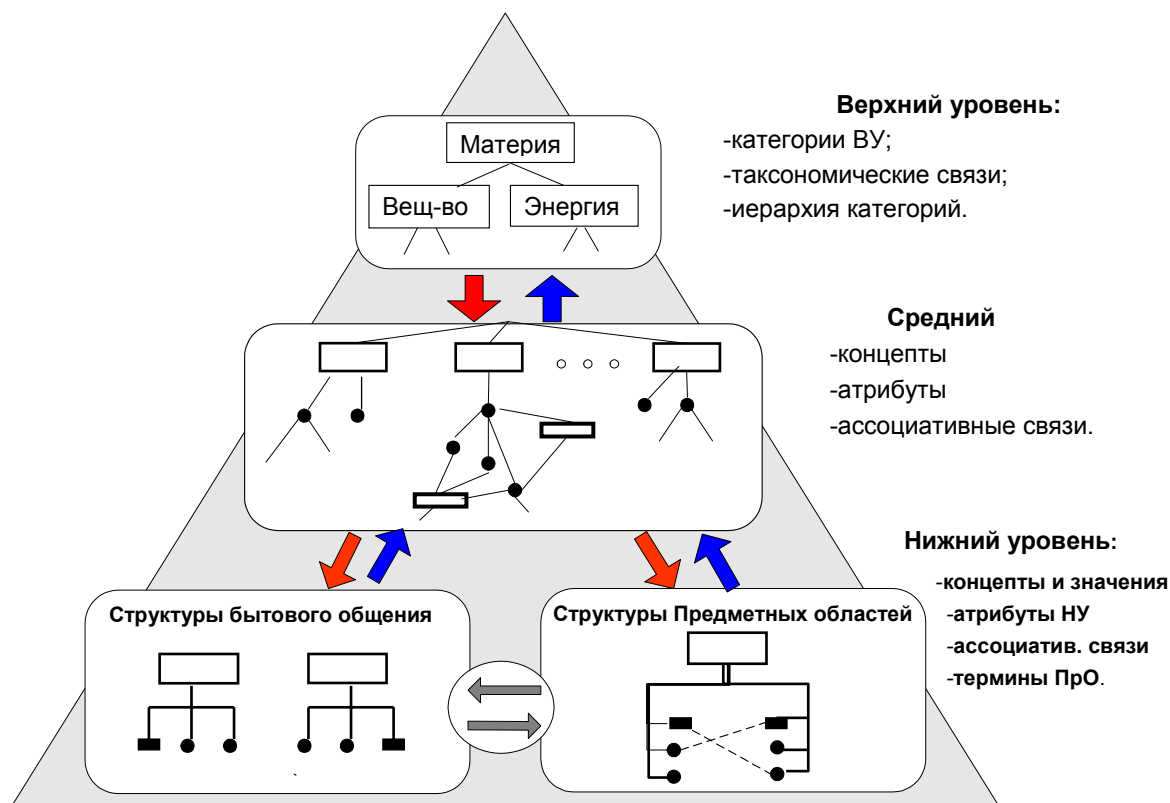


Рис. 1. Иерархическая трёхуровневая онтология – ИО*3. Пирамидальная семантическая сеть, построенная на концептах, отношениях и декларациях.

Коммуникативные когнитивные вычисления

Что мы понимаем под компьютерным обнаружением смысла ЕЯ текстов?

Для того, чтобы связать вместе *текст*, *знания* и *смысл*, нам понадобятся следующие определения:

Смысл текста естественного языка (дискурса) – это конструкция внутри системы общих знаний (онтологии), которая активизируется под воздействием текста и несёт в себе концептуальное изложение вербализованного замысла автора.

Вычисление смысла текста, который выполняет коммуникативную функцию передачи знаний, мы относим к процедурам когнитивного характера, доступным для компьютера. **Извлечение смысла текста ЕЯ** – это информационная технология обработки языковых (текстовых) сообщений, целью которой является выявление в структурированном языковом материале таких когнитивных структур, которые направлены на извлечение смысла сообщения и его понимания на основе знания языка, знаний о языке и использования общих для коммуникантов знаний о мире.

Цель семантической обработки текста ЕЯ – онтологический смысл

Рассмотрим направленный концептуальный граф ИО*3, представляющий собой когнитивную структуру дискретного представления знаний (см. рис. 1). Благодаря пирамидальной иерархии любой путь на графе, ведущий от некоторого *начального концепта* вверх, должен заканчиваться в его вершине – *Материя*; этот путь является подграфом онтологического графа [3].

Подграф онтологического графа, который порождается начальным концептом, называется «смысловой траекторией начального концепта», или «онтологическим смыслом» начального слова.

Результатом онтологического анализа *одного слова* служит его индивидуальная смысловая траектория, а результатом анализа *всего текста* является совокупность смысловых траекторий начальных концептов (если значения начальных концептов совпадают со значимыми словами текста). Таким образом, последовательность значимых слов во времени порождает последовательность смысловых траекторий этих слов – так, что данные траектории совместно образуют смысл текста. **Онтологический смысл текста есть композиция онтологических смыслов значимых слов.**

Процедура получения онтологического смысла является конструктивной, то есть – машинно-интерпретируемой. Она излагается ниже – на конкретном примере онтологического анализа художественного текста с описанием битвы между персами и спартанцами при Фермопилах (V век до н.э.). Из полного рассказа (объёмом две страницы) произвольно выбран один абзац. Этапы обработки заданного фрагмента иллюстрируются в текстовом и графическом виде.

Пример. «Битва спартанцев с персами» (отрывок)

Со стороны персов, темнеющих до горизонта однородной и серой тучей кожаных колпаков и копей, вперёд выступили лучники в кожаных доспехах. Туго натянув тетивы и устремив стрелы почти в зенит, они замерли, словно статуи. Мощные бронзовые мускулы обнажённых рук воинов выпятились буграми предельного нечеловеческого напряжения... Всё замерло.

Тактика врага прояснялась.

Процедура выделения смысла

Первый этап. *Извлечение из текста последовательности значимых слов* (выполняется лингвистическим процессором).

Персы, горизонт, туча, (колпаки), копья, лучники, доспехи, тетивы, стрелы, (зенит), (они), статуи, мускулы, руки, воины, (бугры), напряжение, (всё), замерло = покой, тактика, враг.

В данный список намеренно не попали сказуемые, определения и слова, отсутствующие в экспериментальной онтологии (даны в скобках). Они исключаются, а остальные **16 слов** становятся *начальными именами* и порождают смысловые траектории.

Второй этап. *Построение смысловых траекторий* (список имён сокращён).

1. **Персия** – Государство древнего мира – Организация д.м. – Организация – Общество – Ноосфера – Живое вещество – Вещество – **Материя.**
2. **Горизонт** – Пространство – Природная среда – Природа – Неживое вещество – **Вещество** – Материя.
3. **Туча** – Вода – Жидкость – Состояние – **Неживое Вещество** – Вещество – Материя.
4. **Копьё** – Оружие древнего мира – Война д.м. – Деятельность д.м. – Деятельность – **Общество** – Ноосфера – Живое вещество – Вещество – Материя.
- ...
16. **Враг** – **Воин** – Армия – Война – Деятельность – Общество – Ноосфера – Живое вещество – Вещество – Материя.

Все 16 траекторий являются *полными*, то есть, каждая из них достигает вершины пирамиды знаний. Объединение всех полных траекторий образует на графе структуру, *необходимую и достаточную* для синтеза подграфа онтологического смысла.

Третий этап. Устранение избыточности смысловых траекторий.

В процессе появления новых траекторий с очевидностью проявляется дублирование их конечных отрезков. На каждой траектории выделяются точки пересечения, которые называются *концептами связности*. Дублирующие цепочки траекторий исключаются. Остаётся совокупность *усеченных траекторий* смысла для всех 16-ти начальных имён:

1. *Персия* – Государство древнего мира – Организация д.м. – Организация – Общество – Ноосфера – Живое вещество – Вещество – **Материя.**
2. *Горизонт* – Пространство – Природная среда – Природа – Неживое вещество – Вещество.
3. *Туча* – Вода – Жидкость – Состояние – **Неживое Вещество.**
4. *Копьё* – Оружие др. мира – Война д.м. – Деятельность д.м. – Деятельность – Общество.
5. *Лучник* – Лук – **Оружие д.м.**
6. *Доспехи* – Защита – **Оружие д.м.**
7. *Тетива* – Лук.
8. *Стрела* – Лук.
9. *Статуя* – Скульптура – Искусство – **Деятельность.**
10. *Мускулы* – Тело – Анатомия – Человек – **Общество.**
11. *Руки* – Тело.
12. *Воин* – Армия – Война – **Деятельность.**
13. *Напряжение* – Сила – Энергия – **Материя.**
14. *Покой* – **Состояние.**
15. *Тактика* – Борьба – **Война.**
16. *Враг* – **Воин.**

Результат процедуры выделения смысла

Графическое представление онто-смысла является релевантным тексту и достаточным, чтобы потребитель смог уловить (понять) авторский замысел. Однако, в случае необходимости, онтологический смысл может быть преобразован в литературный. Эта задача должна решаться *лингвистическим транслятором*, который здесь не рассматривается. Для выбранного фрагмента рассказа естественный вариант смысла формулируется следующим образом:

«В древнем мире происходила война, в которой участвовала Персия. Воины применяли доспехи, копья, луки».

После анализа *полного текста*, с адресным привлечением энциклопедических знаний, определений и толкований некоторых слов, входящих в онтологический смысл, литературный смысл рассказа может быть выражен примерно так:

«Греческие воины Спарты проявили отвагу в битве с превосходящим войском персидского царя Ксеркса. Все они погибли, но задержали продвижение армии врага. Их подвиг – пример выполнения воинского долга до конца, любви к Родине».

Заключение

На основе онтологического представления системы знаний и формального определения смысла как подграфа онтологического графа разработана новая информационная технология извлечения смысла текста естественного языка (дискурса) в базе знаний компьютера. Информационную технологию можно использовать для

когнитивних вычислений, в частности – для «машинного понимания» основного содержания семантически-насыщенных текстов, адресных сообщений и документов, в том числе – в профессиональных предметных областях деятельности.

В Системах поддержки принятия решений данная информационная технология может служить *семантическим фильтром* при отборе из Интернет документов – не только релевантных запросу, но имеющих для потребителя прагматический смысл. В **экспертных системах** персональные онтологии, построенные на базе личных знаний, опыта и интуиции позволят подкрепить субъективные экспертные оценки объективно-вычисленным смыслом, то есть – *обоснованием принятия тех или иных решений*.

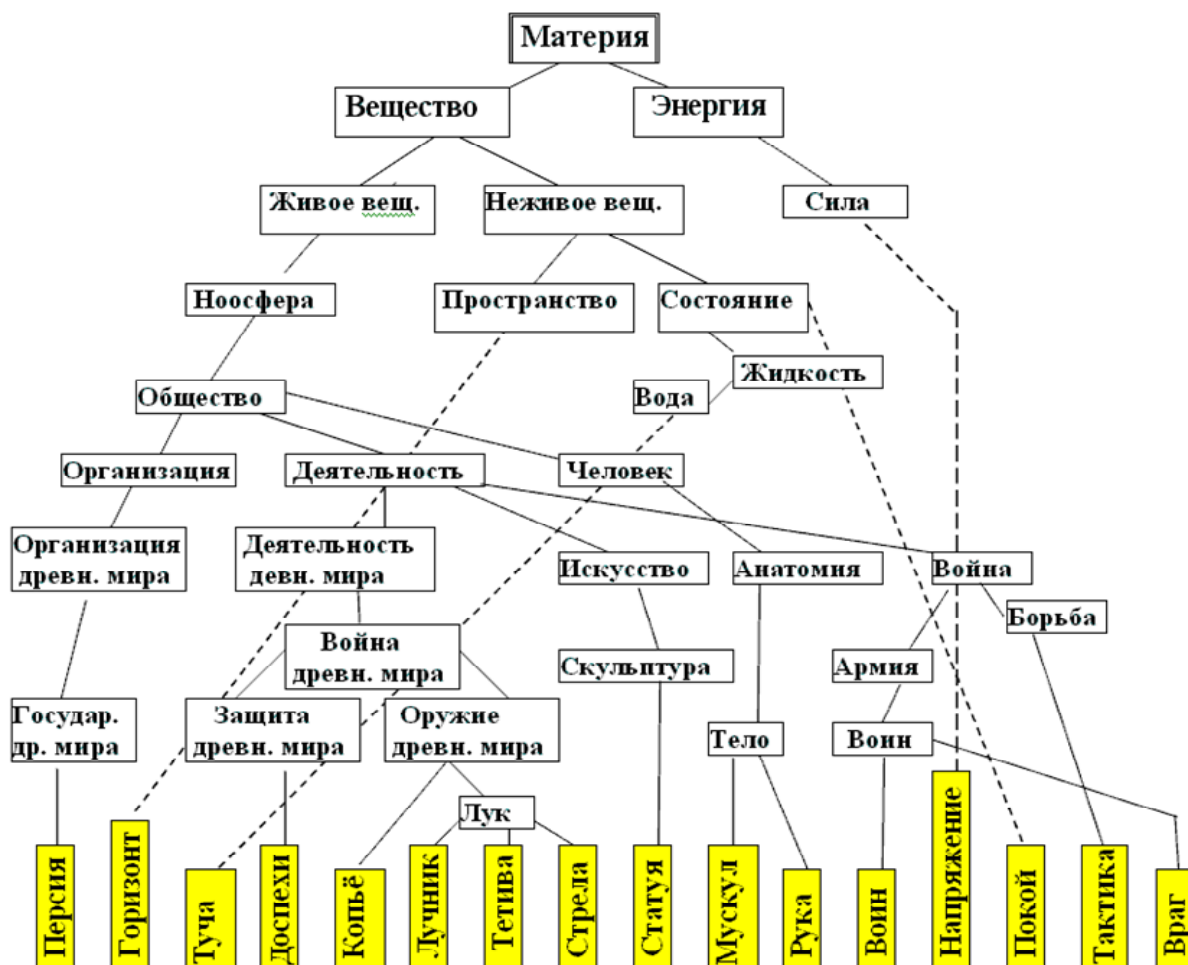


Рис. 2. Смысловые траектории, образующие в совокупности онтологический смысл фрагмента рассказа «Битва спартанцев»

Литература

1. Демьянков В.З. Основы теории интерпретации и её приложения в вычислительной лингвистике. – М.: Изд-во Моск. Ун-та, 1985. – 76 с.
2. Таран Т.А., Зубов Д.А. Искусственный интеллект. Теория и приложения. Научное пособие. – Луганск: Изд. ВНУ им. В. Даля, 2006. – 240 с.
3. В. Гладун, В. Величко, Л. Святогор. Структурирование онтологии ассоциаций для конспектирования естественно-языковых текстов // Advanced Research in Artificial Intelligence. International Book Series, Number 2. Supplement to the International Journal “Information Technologies & Knowledge” Volume 2 / 2008. – ITHEA, Sofia, 2008. – p. 153.