References

1 Bukhgoltc N.N. *Basic course of theoretical mechanics,* part 1, Moscow: Science, 1972, p. 468.

2 Levitckiy N.I. *Theory of mechanisms and machines*, Moscow: Science, 1990, p. 592.

3 Dzholdasbekov U.A., Biyarov T.N. *Stability and stabilization of movement of mechanisms and cars* / Preprint. № 3 IA RK, Almaty, 1993, p. 82.

УДК 519.7

А.Т.Абдрахманов, Р.Р.Мусабаев, Н.Тасболатулы

Институт проблем информатики и управления, Алматы (E-mail: ata61@mail.ru)

Семантические сети для смыслового анализа текстов

Рассмотрены задачи семантического анализа и обработки текстов, являющиеся наиболее актуальными проблемами компьютерной науки последних десятилетий. В настоящее время существуют различные базы данных для семантического анализа текстов. Одними из наиболее широко применяемых направлений являются WordNet, EuroWordNet, BalkaNet и RussNet. Семантическая обработка текста осуществляется в три этапа: морфологический, синтаксический и семантический анализы. В статье доказана необходимость применения алгоритма семантического анализа для дальнейшего развития методов поиска текстовой информации, а также приведен обзор технологий семантических сетей для смыслового анализа текстовых данных.

Ключевые слова: семантическая сеть, WordNet, смысловой анализ, морфологический, синтаксический и семантический анализы, онтология, синсет, фрейм.

1 Семантическая сеть

Целью данной статьи является обзор технологий семантических сетей для смыслового анализа текстовых данных. На сегодняшний день существуют различные базы данных для смыслового анализа текстов. Одними из самых широко применяемых разработок являются WordNet [1], EuroWordNet [2], BalkaNet [3] и RussNet [4]. Семантическая сеть — информационная модель предметной области, имеющая вид ориентированного графа, вершины которого соответствуют объектам предметной области, а дуги задают отношения между ними. Объектами могут быть понятия, события, свойства, процессы [5].

Значительный вклад в изучение и исследование проблемы смыслового анализа текстов внесли зарубежные и отечественные ученые, в частности, George A.Miller, Christiane Fellbaum, R.Beckwith, D.Gross, K.Miller, С.Э.Фалман, Т.Бернерс-Ли, Г.Морис, А.Н.Баранов, Д.О.Добровольский, М.В.Дибривный, Д.Е.Шуклин, А.А.Шарипбаев, Б.Ш.Разахова, А.К.Жубанов и другие [6–10].

Одним из успешных проектов смыслового анализа текстов является WordNet. Работа над словарем WordNeta английского языка начата в Принстонском университете (США) в начале 80-х годов и продолжается до настоящего момента. Сейчас доступна версия 3.0 этого словаря. Существующая версия WordNet (см. табл.) охватывает общеупотребительную лексику современного английского языка. Широкое распространение этот словарь получил благодаря его свободной доступности для научных и исследовательских целей.

Таблица

Статистика WordNet 3.0

Части речи	Число уникальных строк	Синсеты	Всего пар значений
Существительные	117798	82115	146312
Глаголы	11529	13767	25047
Прилагательные	21479	18156	30002
Наречия	4481	3621	5580
Общее число	155287	117659	206941

В период с марта 1996 по сентябрь 1999 года при финансировании Европейской комиссии был создан многоязычный вариант WordNet — EuroWordNet. Эта лексическая система объединила в себе WordNet словари английского, датского, испанского, итальянского, немецкого, французского, чешского и эстонского языков, за основу был взят Принстонский WordNet. В 2004 году завершена работа над проектом BalkaNet, объединяющим греческий, болгарский, турецкий, чешский, французский, румынский и сербский языки. Все национальные версии WordNet связаны с исходным WordNet и между собой через специальный ILI–индекс (*Interlingualindex* — межъязыковой индекс). Словари EuroWordNet являются коммерческим продуктом [11].

Русская версия WordNet, EuroWordNet и других подобных ресурсов является RussNet. Проект RussNet унаследовал основные особенности указанных выше ресурсов для русского языка. Его разработка ведется кафедрой математической лингвистики филологического факультета Санкт-Петербургского государственного университета с 1999 года. Исследовательской группой руководит И.В.Азарова. До настоящего времени в проекте приняли участие 57 лингвистов и программистов, основная группа состоит из 8 человек. При разработке аналога WordNet, ориентированного на русский язык, было выявлено свыше 120 тыс. лексических единиц, которые описывают лексику русского языка, и сопоставлено по числу лексических единиц с английской версией. Для этого используются морфологический анализатор, лексические ресурсы, словари, свободно распространяемые в Интернете, и ряд печатных изданий. Например:

– интеграция с другими лексическими системами на основе использования технологии SW (SemanticWeb);
– автоматизированное построение межъязыкового индекса, определяющего соответствие между синсетами, на основе использования электронных версий словарей издательства OxfordPress, ряда доступных в Интернете англо-русских и русско-английских словарей, WordNet-Domains.

На сегодняшний день RussNet включает 55397 существительных, образующих 71729 синсетов; 34400 глаголов, образующих 44998 синсетов; 25315 прилагательных, образующих 33571 синсет; 10071 наречие, образующее 9716 синсетов [12].

С 2012 года в ИПИУ МОН РК (Институт проблем информатики и управления Министерства образования и науки Республики Казахстан) в рамках лаборатории «Анализ и моделирование информационных процессов» по проекту «Разработка системы смыслового анализа поиска текстов нового поколения, ориентированного на казахский язык» ведется работа по созданию KazWordNet.

2 Использование семантической сети

Семантические сети вначале использовались для представления смысла выражений естественного языка человека, откуда и появилось название этого класса сетей. Теперь же они используются в качестве структуры, пригодной для представления информации общего вида, — узлы представляют некоторые концепты (понятия), а связи — отношения между концептами. На рисунке 1 показан пример семантической сети.

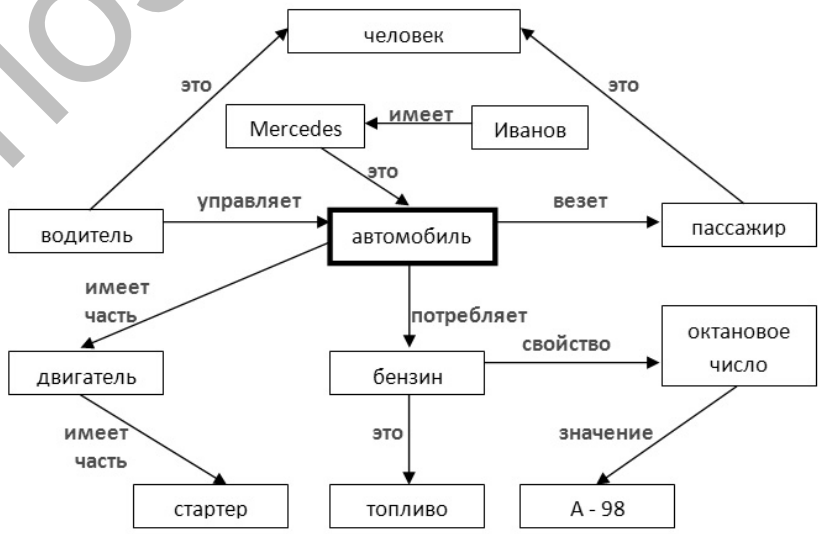

Рисунок 1. Пример семантической сети

В семантических сетях декларативные и процедурные знания не разделены, следовательно, база знаний не отделена от механизма вывода. Процедура логического вывода обычно представляет совокупность процедур обработки сети. Семантические сети получили широкое применение в системах распознавания речи и интеллектуальных системах.

Семантические сети используются в системах информационного поиска (informationretrieval), вопросно-ответных системах (Q&A systems), в системах машинного перевода (machinetranslation) и при решении задачи определения значения слов (WSD — word-sensedisambiguation) [13].

Семантическая обработка текста выполняется в три этапа: морфологический, синтаксический и собственно семантический анализ (рис. 2). Каждый этап выполняет отдельный анализатор со своими входными и выходными данными и собственными настройками.

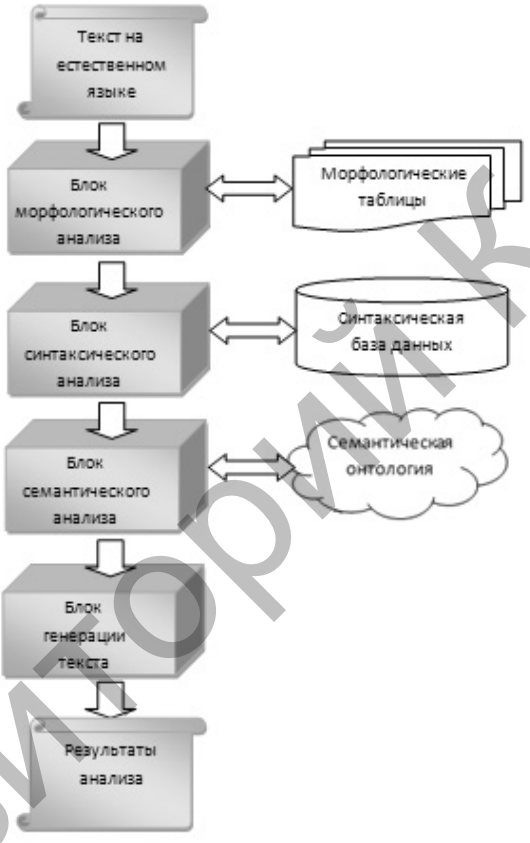

Рисунок 2. Схема семантического анализа текстов

Как видно из рисунка 2, центральное место при такой модели поиска информации в Интернете занимает семантическая онтология. Однако процесс создания семантической онтологии сложный процесс. Информационные семантические онтологии состоят из экземпляров, понятий, атрибутов и отношений между ними. Для создания онтологии необходимо создать словарь терминов — глоссарий, объединить термины общими связями (синсетами) и затем наложить ограничения на эти связи, как показано на рисунке 3.

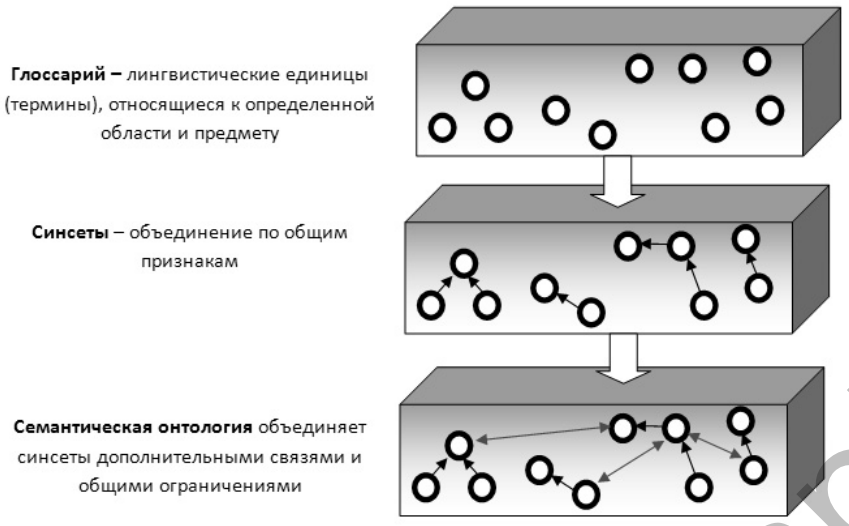

Рисунок 3. Процесс создания семантической онтологии

Для построения онтологий необходимо разработать языки их представления. При этом могут быть использованы такие специализированные языки, как ResourceDescriptionFramework (RDF), WebOntologyLanguage (OWL) и т.д. Онтологии могут использовать различные модели представления знаний, такие как логика предикатов (Firstorderlogics — FOL), дескриптивная логика, фреймовые модели (Frames), концептуальные графы и т.п. Для создания онтологий могут использоваться различные редакторы, которые в свою очередь могут поддерживать различные форматы представления данных (языки), основанные на различных формализмах (логиках, моделях представления данных). Ключевым моментом в проектировании онтологии является выбор соответствующего языка спецификации онтологий (Ontologyspecificationlanguage) и редактора для работы с ней.

Ниже рассмотрен пример применения семантической сети для информационного поиска данных в Интернете.

На рисунке 4 показана схема семантического поиска информации в Интернете. Пользователь вводит запрос, который подвергается лингвистическому анализу, расширяется за счет использования синонимов, затем преобразовывается в ключевые слова и отправляется поисковой машине. Поисковая машина возвращает найденные документы, они также подвергаются лингвистическому разбору и формируются семантические образы документов. Образы документов сравниваются с образом запроса, делается вывод о релевантности каждого из документов и результаты анализа (документы, которые были признаны релевантными) предоставляются пользователю [14].

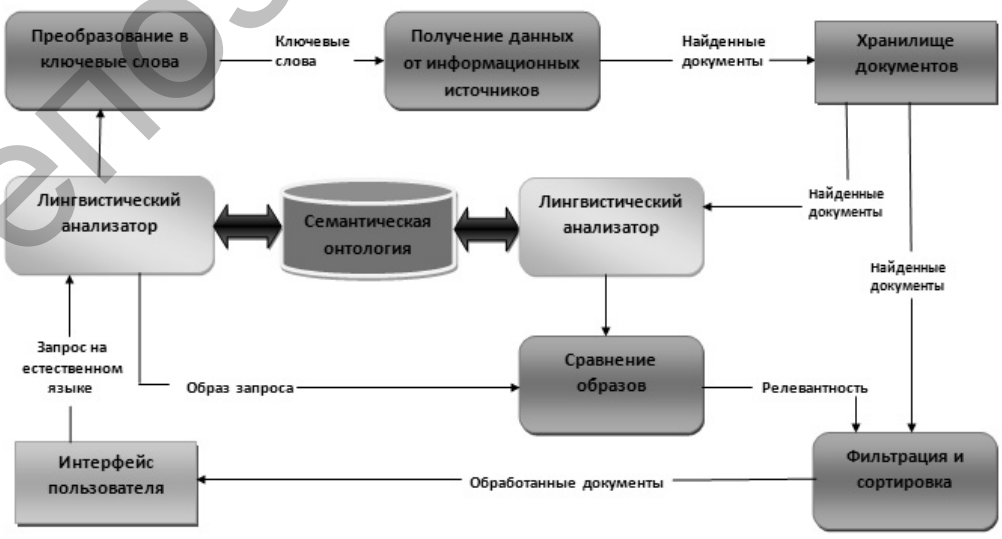

Рисунок 4. Диаграмма потоков данных при поиске

Заключение

Объемы текстовой информации, хранимой в сети Интернет, постоянно увеличиваются. Возникают проблемы с эффективным поиском информации в сети Интернет. В существующих системах поиска текстовой информации широко используются методы поиска по совпадению ключевых слов. Данные методы не всегда показывают приемлемые результаты поиска.

В данной статье было обосновано применение алгоритма семантического анализа для дальнейшего развития методов поиска текстовой информации. Решение задачи построения систем смыслового поиска текстов является комплексной проблемой. По всей видимости, в последующее десятилетие к решению данной проблемы будут приложено значительное усилие специалистов в области интеллектуального поиска информации.

Список литературы

1 *Fellbaum C. (ed.)* WordNet: An Electronic Lexical Database // MIT Press. — 1998.

2 *Vossen P.* Building a multilingual database with WordNets for several European languages // http://www.illc.uva.nl/EuroWordNet/

3 *Horák A., Smrž P.* VisDic – WordNet Browsing and Editing Tool / P.Sojka, K.Pala, P.Smrž, C.Fellbaum, P.Vossen (Eds.): GWC 2004, Proceedings. — P. 136–141. Masaryk University, Brno, 2003.

4 *Азарова И.В., Синопальникова А.А., Яворская М.В.* Принципы построения WordNet-тезауруса RussNet // Тр. конф. «Диалог-2004» // http://www.dialog-21.ru/Archive/2004/Sinopalnikova.htm

5 *Roussopoulos N. D.* A semantic NetWork model of data bases // TR. — № 104. — Department of Computer Science. — University of Toronto. — 1976.

6 *George A. Miller, Christiane Fellbaum*. Semantic NetWorks of English // Cognitive Science Laboratory. — Princeton University. Princeton (USA), 2002.

7 *Баранов А.Н., Добровольский Д.О.* Постулаты когнитивной семантики // Изв. АН. РФ Сер. литературы и языка. — 1997. — Т. 56. — № 1. — С. 11–21.

8 *Шуклин Д.Е.* Структура семантической нейронной сети, реализующей морфологический и синтаксический разбор текста // Кибернетика и системный анализ. — Киев: Изд-во Ин-та кибернетики НАН Украины, 2001. — № 5. — С. 172–179.

9 *Шарипбаев А.А., Разахова Б.Ш., Кабенов Д.И.* Интеллектуальная система оценки знаний, основанная на семантических сетях. — Астана: ЕНУ им. Л.Н. Гумилева, 2011.

10 *Жубанов А.К.* Основные принципы формализации содержания казахского текста: Дис. ... д-ра филол. наук. — Алматы, 2002.

11 *Сухоногов А.М., Яблонский С.А.* Разработка русского WordNet // Тр. 6-й Всеросс. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции». — RCDL 2004, Пущино, Россия, 2004.

12 *Сухоногов А.М., Яблонский С.А.* Лексические онтологии WordNet в технологиях SemanticWeb // Междунар. журн. «Программные продукты и системы»; «Роспечать». — 2009. — № 4.

13 *Мозговой М.В.* Машинный семантический анализ русского языка и его применение: Дис. ... канд. физ.-мат. наук. — СПб., 2006.

14 *Марченко О.О.* Моделювання семантичного контексту при аналізі текстів на природній мові // Вісн. Київс. ун-ту. Сер. фіз.-мат. науки. — 2006. — № 3. — С. 230–235.

А.Т.Әбдірахманов, Р.Р.Мұсабаев, Н.Тасболатұлы

Мәтіндердің мағыналы талдауы үшін семантикалық желілер

Мәтінді мағыналық тұрғыда талдау және өңдеу мәселесі компьютерлік ғылым саласының соңғы онжылдықтардағы ең өзекті мәселелерінің бірі болып табылады. Қазіргі кезде мәтінді мағыналық жағынан біріктірген ірі деректер қоры жазылған, олар: WordNet, EuroWordNet, BalkaNet, RussNet. Лексикалық тексеруден өткен мәтінді семантикалық талдау үш сатыдан тұрады: морфологиялық, синтаксистік және семантикалық талдау. Мақалада мәтіндік ақпаратты іздестіру жүйелерінде қолданылатын қарапайым семантикалық талдау алгоритмі қарастырылды.

A.T.Abdrakhmanov, R.R.Musabayev, N.Tasbolatuly

Semantic networks for the semantic analysis of texts

We considered the task of the semantic analysis and processing of texts which is one of the most actual problems of a computer science of the last decades. There are various databases for the semantic analysis of texts at present. One of the most widely applied development are WordNet, EuroWordNet, BalkaNet and RussNet. Semantic text processing is carried out in three stages: morphological, syntactic and actually semantic analysis. In this article application of algorithm of the semantic analysis for further development of methods of search of text information was carried out.

References

1 Fellbaum C. *(ed.) WordNet: An Electronic Lexical Database*, MIT Press, 1998.

2 Vossen P. *Building a multilingual database with WordNets for several European languages*. http://www.illc.uva.nl/EuroWordNet/

3 Horák A., Smrž P. *VisDic — Wordnet Browsing and Editing Tool P.Sojka, K.Pala, P.Smrž, C.Fellbaum, P.Vossen (Eds.): GWC 2004, Proceedings*, p. 136–141, Masaryk University, Brno, 2003.

4 Azarov I.V., Sinopalnikova A.A., Yavorskaya M.V. *Principles of creation of wordnet-thesaurus RussNet // «Dialogue – 2004» conference works*, http://www.dialog-21.ru/Archive/2004/Sinopalnikova.htm.

5 Roussopoulos N.D. *A semantic NetWork model of data bases // TR № 104,* Department of Computer Science, University of Toronto, 1976.

6 George A.Miller, Christiane Fellbaum. *Semantic NetWorks of English //* Cognitive Science Laboratory, Princeton University, Princeton, U.S.A., 2002.

7 Baranov A.N., Dobrovolsky D.O. *Postulates of cognitive semantics //* AN news. Literature and language series, 1997, vol. 56, № 1, p. 11–21.

8 Shuklin D.E. *Structure of the semantic neural NetWork realizing morphological and syntactic analysis of the text* // Cybernetics and the system analysis. Kiev: Publ. house of Institute of cybernetics of NAN of Ukraine, 2001, № 5, p. 172–179.

9 Sharipbayev A.A., Razakhova B.Sh., Kabenov D.I. *Intellectual system of an assessment of the knowledge, based on semantic networks* // The Eurasian National University of L.N.Gumilev, Astana, 2011.

10 Zhubanov A.K. *Basic principles of formalization contents of the Kazakh text:* The dissertation on competition of a scientific degree of the Doctor of Philology, Almaty, 2002.

11 Sukhonogov A.M., Yablonsky S.A. *Russian WordNet development* // Works of the 6th All-Russia scientific conference «Electronic libraries: perspective methods and technologies, electronic collections». — RCDL, 2004, Pushchino, Russia, 2004.

12 Sukhonogov A.M., Yablonsky S.A. *Lexical ontologies of wordnet in the semantic web research* // International magazine «Software products and systems», «Rospechat». — № 4. — 2009.

13 Mozgovoy M.V. *Machine semantic analysis of Russian and its application* // The dissertation on competition of a scientific degree of the candidate of physical and mathematical sciences, Sankt-Petersburg, 2006.

14 Marchenko O.O. *Modeling of a semantic context in the analysis of texts on natural language* // The Messenger of the Kiev university, Physical and mathematical series. nauka, 2006, № 3, p. 230–235.