ISSN 2306-1561
Automation and Control in Technical Systems (ACTS)
2014, No 4, pp. 109-118.
DOI: 10.12731/2306-1561-2014-4-11

Methodology of Computer-Aided Learning of Mivar Knowledge Base for Computer Understanding of Russian Language

Peterson Anastasia Olegovna

Russian Federation, Undergraduate Student, Department of «Automated Control Systems».

State Technical University – MADI, 125319, Russian Federation, Moscow, Leningradsky prospekt, 64. Tel.: +7 (499) 151-64-12. http://www.madi.ru

vasileva5193@gmail.com

Abstract. One of the main aims of artificial intelligence is natural language understanding. In this article described some methods, which can allow to reach this goal. The possibility of using mivar technologies in modeling of text meaning understanding has been proved theoretically. “Mivar-Text” is a software complex for text meaning understanding. The necessity of automatic filling of it knowledge base has been explained. In this article has been proposed new methods and algorithms of automatic knowledge base learning and. New module of knowledge base filling has been created based on them and integrated into "Mivar-Text". The algorithms for automated parsing of words compatibility dictionary and automated definition of words compatibility has been implemented in the form of program. Mivar technology has proven its advantages in the understanding of natural language: the real-time processing of huge volumes of data and rules. Given the advances of mivar technologies in the accumulation and processing of information, the solution of the problem of computer-aided knowledge base learning will move to a new level of automatic processing of texts (APT) based on logical processing large arrays of data and context.

Keywords: mivar, mivar net, mivar technologies, production rules, natural language, artificial intelligence, automatic text processing, Russian language.

ISSN 2306-1561
Автоматизация и управление в технических системах (АУТС)
2014. – №4. – С. 109-118.
DOI: 10.12731/2306-1561-2014-4-11

УДК 004.9

Разработка методики автоматизированного обучения миварной базы знаний в целях понимания компьютерами естественного русского языка

Петерсон Анастасия Олеговна
Российская Федерация, магистрант кафедры «Автоматизированные системы управления».
ФГБОУ ВПО «Московский автомобильно-дорожный государственный технический университет (МАДИ)», 125319, Российская Федерация, г. Москва, Ленинградский проспект, д.64, Тел.: +7 (499) 151-64-12, http://www.madi.ru
vasileva5193@gmail.com

Аннотация. Рассмотрены методы реализации одной из основных целей создания искусственного интеллекта: понимание компьютерами естественного языка. Теоретически обоснована целесообразность применения миварного подхода к моделированию понимания смысла текстов. Обоснована необходимость автоматизированного обучения миварной базы знаний в программном комплексе понимания текстов «МИВАР-Текст». В работе предложены новые: методика и алгоритмы автоматизированного обучения миварной базы знаний; создана подсистема «Автоматизированное заполнение миварной базы знаний» в программном комплексе понимания естественного русского языка «МИВАР-Текст». Разработаны и реализованы в виде программ: алгоритм автоматизированного разбора словаря сочетаемостей и алгоритм автоматизированной расстановки сочетаемостей. Миварные технологии доказали на практике свои преимущества в области понимания естественного языка: обработка в реальном времени сверхбольших объемов данных и правил. Учитывая достижения миварных технологий в накоплении и обработке информации, решение задачи автоматизированного обучения баз знаний позволит перейти на новый качественный уровень автоматизированной обработки текстов (АОТ) на основе логической обработки больших массивов данных и учета контекста.

Ключевые слова: мивар, миварные сети, продукции, естественный язык, искусственный интеллект, миварные технологии, интеллектуальные системы, русский язык.

1. Введение

Одной из основных целей создания искусственного интеллекта является понимание компьютерами естественного языка, т.е. способность программ понимать человеческий язык и строить фразы на нем. В настоящее время, ученые понимают, что такая способность должна учитывать «контекст», т.е. базируется на обширном фоновом знании о предмете беседы и идиомах, используемых в этой области, так же, как и на способности применять общее контекстуальное знание для понимания недомолвок и неясностей, присущих естественной человеческой речи [1 – 5].

Как известно, естественный язык — это язык, используемый для общения людей (в отличие от формальных языков и других типов знаковых систем, также называемых языками в семиотике) и не созданный целенаправленно (в отличие от искусственных языков) [7]. Словарь и грамматические правила естественного языка определяются практикой применения и не всегда бывают формально зафиксированы. В связи с этим возникают проблемы понимания естественного языка при создании искусственного интеллекта. Задача сбора и организации фонового знания, чтобы его можно было применить к осмыслению языка, составляет значительную проблему в автоматизации понимания естественного языка.

Для ее решения исследователи разработали множество методов структурирования семантических значений, используемых повсеместно в искусственном интеллекте [1 – 8]. Актуальность темы работы обусловлена тем, что до сих пор проблема понимания компьютерами естественного языка не решена, а потребность в ее решении только увеличивается.

2. Обзор задач и работ по пониманию текстов

Основная часть текущих работ в области понимания естественных языков направлена на поиск формализмов представления, которые должны быть достаточно общими, чтобы применяться в широком круге приложений и уметь адаптироваться к специфичной структуре заданной области.

Множество разнообразных методик (большинство из которых являются развитием или модификацией семантических сетей) исследуются с этой целью и используются при разработке программ, способных понимать естественный язык в ограниченных, но достаточно интересных предметных областях [1 – 9]. Общим недостатком таких подходов является недостаточное внимание к накоплению и быстрой обработке информации о всей предметной области, т.е. контекстов. Общепризнано, что без учета контекста нельзя решить задачу понимания естественного языка [1 – 31].

В настоящее время ведутся работы по созданию программного комплекса понимания текстов «МИВАР-Текст». Этот комплекс должен эволюционно обучаться на основе миварных баз данных и правил (МБДП), которые для краткости далее называются «миварные базы знаний». Известны способы ручного обучения миварных баз знаний (БЗ), которые описаны в научных работах [7].

Практическая реализация «МИВАР-Текст» показала, что ручное обучение является не достаточно эффективным и требует огромных затрат времени и человеческих ресурсов. Конечно, на начальных этапах ручное обучение является необходимым, например при описании предметных областей из 150 – 300 слов-терминов. Такое количество слов соответствует словарному запасу трехлетнему ребенку. Однако, в процессе обучения у ребенка появляется способность к чтению и самостоятельному обучению на основе известного минимума слов. В нашем проекте постепенное обучение системы «МИВАР-Текст» аналогично обучению ребенка и с определенного момента возникла необходимость автоматизированного обучения для достижения словарного запаса более 3000 слов. Практика показала, что к этому моменту система уже начинает понимать адаптированные тексты и способна к эволюционному наращиванию БЗ [7].

Конечной целью создания «МИВАР-Текст» является способность к самостоятельному автоматическому обучению системы на основе получения и автоматической обработки обучающих (эталонных, проверенных) текстов, например: словарей, учебников и энциклопедий. Однако, в настоящий момент доступны только отсканированные версии таких словарей, содержащие множество ошибок из-за недостаточно хорошей работы автоматических распознавателей текстов и плохого качества исходного материала.

Текст из таких словарей можно вводить в миварные БЗ только после автоматизированной проверки и специальной обработки, требующей участия человека-оператора. Подчеркнем, что такое участие человека не является «ручным обучением» и даже в самых сложных случаях позволяет повысить производительность обучения миварной БЗ в сотни раз, по сравнению с традиционным ручным обучением. Отметим, что задачи автоматизированного обучения миварных баз знаний были недостаточно проработаны в предыдущих исследованиях [1 – 31].

Таким образом, существует актуальная сложная научная проблема: необходимо автоматизировать процесс обучения миварных баз знаний, основанный на «чтении отсканированных словарей» [7].

3. Модели и методы решения задач. Миварный подход к пониманию смысла текста

В качестве моделей решения задач применяется миварное информационное пространство, подробно описанное в работах [7 – 31]. Миварное информационное пространство базируется на эволюционных многмоерных базах данных и правил, которые позволяют накапливать любые данные о любых контекстах и правилах работы со словоформами русского языка.

В качестве методов решения задач применяются метод накопления информации в виде миварных сетей и метод логико-вычислительной обработки с линейной вычислительной сложностью, которые более подробно изложены в литературе [4 – 31].

Миварный подход к пониманию смысла текста базируется на математическом отображении частей речи русского языка (существительное, глагол, прилагательное, местоимение, причастие, деепричастие, наречие, числительное и др.) в основные

понятия концептуальной модели миварного информационного пространства: "вещь, свойство и отношение".

Конечно, синтаксис используется, но его роль сведена только к выделению взаимосвязи слов в основных словосочетаниях. Используется специальный словарь словоформ с набором морфологических признаков, которые хранятся в базе данных. Кроме морфологических признаков в БД накапливаются другие служебные признаки.

Миварный подход включает технологии накопления данных и обработки информации в едином миварном информационном пространстве. Для понимания языка надо собрать и поддерживать в актуальном состоянии огромную базу данных фактов и такое же большое количество правил, которые позволяют выявлять нюансы смысла разных понятий в различных ситуациях [4 – 31].

Целью работы является разработка методики автоматизированного заполнения БЗ, необходимой для работы с текстами. Необходимо уменьшить вмешательство человека и максимально автоматизировать процесс заполнения базы знаний.

База знаний (БЗ) – это ядро для построения миварной системы анализа текстов и ответов на вопросы, в которой хранятся данные и правила.

Для достижения поставленной цели, решены следующие задачи проекта:

- разработана методика автоматизированного обучения миварной базы знаний;
- разработан алгоритм автоматизированного разбора словаря сочетаемостей;
- разработан алгоритм автоматизированной расстановки сочетаемостей;
- создана подсистема автоматизированного обучения для программного комплекса «МИВАР-Текст».

4. Результаты решения задачи создания подсистемы автоматизированного обучения миварных баз знаний

Подсистема автоматизированного заполнения миварной базы данных необходима для ускорения процесса обучения баз знаний и для того, чтобы избежать неоднозначности, зачастую возникающей при понимании текстов на русском языке.

В настоящее время в НИИ МИВАР ведутся работы по созданию программного комплекса понимания текстов на естественном русском языке «МИВАР-Текст», которая включает в себя подсистему «Обучение».

Подсистема «Обучение» делится на блоки:

1. Формирование сетей понятий(концептов);
2. Автоматизированное заполнение БЗ;
3. Обучение ВЛ.

На рисунке 1 графически показана структура подсистемы «Автоматизированное заполнение миварной базы знаний». В качестве входных данных используются общеизвестные словари (например, Толковый словарь Ожегова) отсканированные при помощи программы Fine Reader.

Отметим, что отсканированные тексты всегда содержат большое количество ошибок, которые можно исправлять в ручном или автоматизированном порядке. Если использовать заранее правильно написанные тексты в электронном виде, то можно будет автоматически вводить информацию из них в базы знаний.

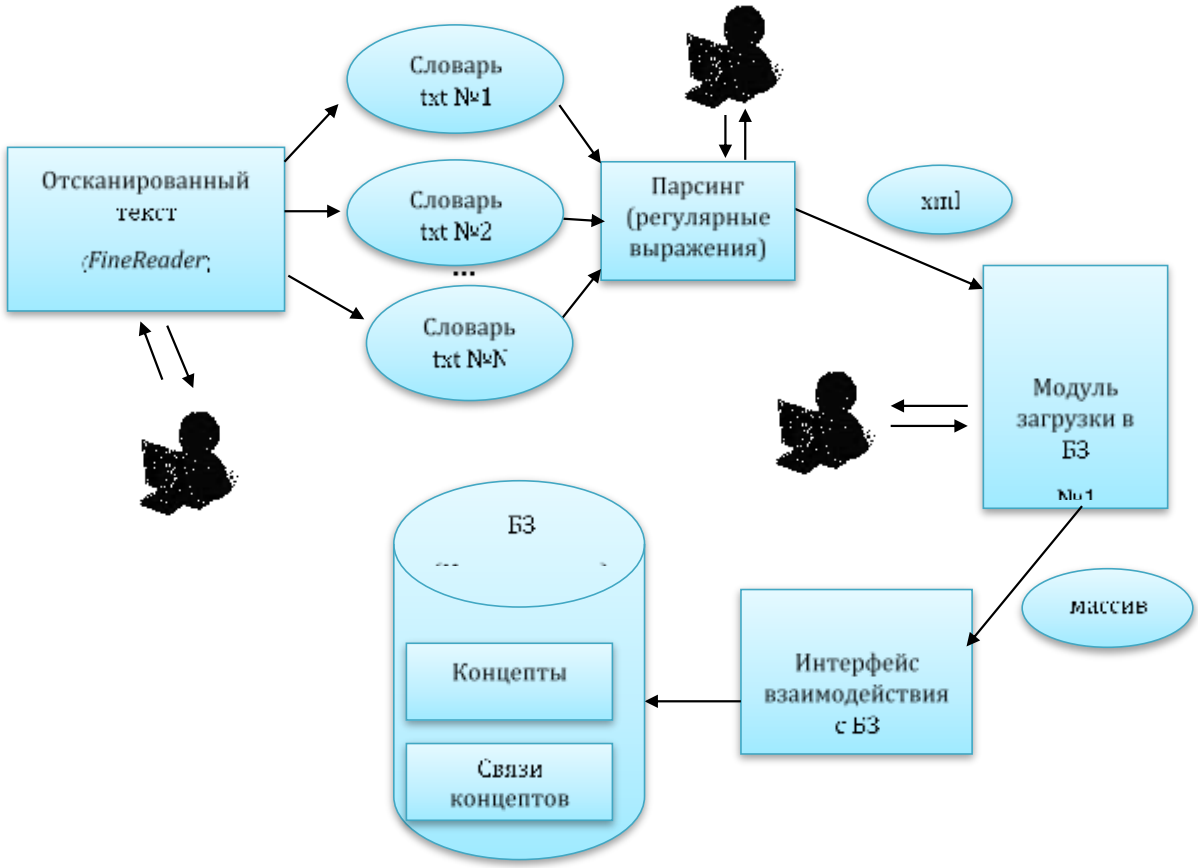

Рисунок 1 – Структура блока автоматизации заполнения миварной БЗ

Данная подсистема необходима для ускорения процесса обучения баз знаний и для того, чтобы избежать неоднозначности, зачастую возникающей при понимании текстов на русском языке.

Этапы автоматизированного заполнения миварной базы знаний:

1) автоматизированный разбор словаря сочетаемостей.
   Заполнение таблицы с концептами с помощью регулярных выражений, обучение БЗ. Осложняющим фактором является то, что на входе отсканированный текст с ошибками, требующий ручной обработки.
2) определение признака сочетаемостей концептов.
   Расстановкой связей между основными концептами из словарной статьи с концептами сочетаемых слов.

Назначение подсистемы: автоматизированное обучение БЗ при помощи выделения СИС (словосочетания, имеющие смысл) и получение отдельных концептов из словаря. Процесс обучения – составление миварной сети (двудольный ориентированный граф), где концепты являются вершинами, а связи между ними – ребра графа [7].

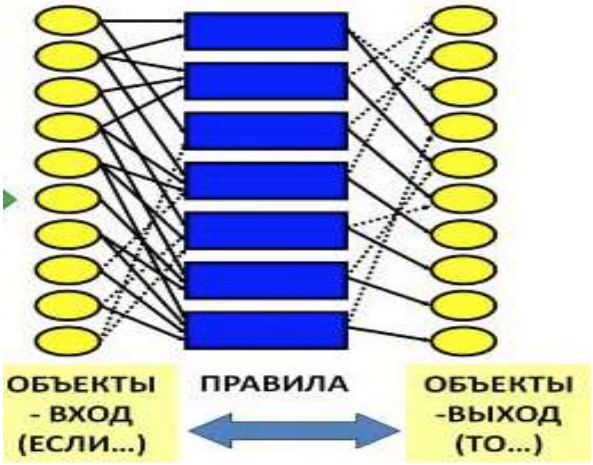

Рисунок 2 – Двудольный ориентированный граф

Создание миварной сети – это параллельное выявление миварных правил и занесение их в базу данных. Постепенно они связываются в полноценную миварную сеть довольно большого объема. Важно, что это происходит параллельно и достаточно быстро [7].

В созданной подсистеме автоматизированного обучения миварной базы знаний реализованы следующие возможности:

1) ручной вставки текстов в окна ввода;
2) проверки и редактирования текста;
3) выполнения разбора(парсинг) словарных статей;
4) создания xml-файла;
5) модуль загрузки должен преобразовывать полученные xml-файлы в массивы миварной базы знаний.

Используя методы автоматизированного обучения в системе «МИВАР-текст», за 3 месяца работы удалось добиться следующих результатов: содержание семантического ядра системы – около 1500 концептов (вершин графа), провязанных между собой в 7 системных отношениях, а также «свободных» (не имеющих провязку) 2000 понятий. Всего связей (ребер графа) – более 200 тысяч.

Системные отношения, используемые (понимаемые) в «МИВАР-Текст»: словосочетания, имеющие смысл (СИС), общее-частное (ОЧ), часть-целое (ЧЦ), находится (то, что должно находится вместе где-либо), свойство, антоним, значение, свойство.

Миварные технологии доказали на практике свои преимущества в области понимания естественного языка: обработка в реальном времени сверхбольших объемов данных и правил. Отметим, что цель работы достигнута и все задачи успешно решены.

Заключение

В настоящее время ведутся работы по созданию программного комплекса понимания текстов «МИВАР-Текст», частью которого является подсистема «Автоматизированное заполнение миварной базы знаний».

Практическая реализация «МИВАР-Текст» показала, что ручное обучение является не достаточно эффективным и требует огромных затрат времени и человеческих ресурсов.

За счет автоматизации обработки и загрузки обучающих текстов из отсканированных словарей, был значительно сокращен срок процесса обучения миварной базы знаний (БЗ), что говорит о достижении цель работы.

Используя методы автоматизированного обучения в системе «МИВАР-текст», за 3 месяца удалось создать миварную сеть, включающую более 1500 концептов (вершин графа) и более 200 тысяч ребер графа. При этом, на такой размерности графа система работает в реальном времени. Миварные технологии доказали на практике свои преимущества в области понимания естественного языка: обработка в реальном времени сверхбольших объемов данных и правил.

Учитывая достижения миварных технологий в накоплении и обработке информации, решение задачи автоматизированного обучения баз знаний позволит перейти на новый качественный уровень автоматизированной обработки текстов на основе логической обработки больших массивов данных и учета контекста.

Список информационных источников

[1] Люгер Дж.Ф. Искусственный интеллект: стратегии и методы решения сложных проблем, 4-е издание: пер. с англ. – М.: Издательский дом "Вильямс", 2005.

[2] Лорьер Ж.-Л. Системы искусственного интеллекта: Пер. с франц. – М.: Мир, 1991. – 568 с.

[3] Белоус Е.С., Кудинов В.А., Желнин М. Э. Современные модели представления знаний в обучающих системах // Ученые записки. Электронный научный журнал Курского государственного университета – 2010, № 1, С. 9-14.

[4] Варламов О.О. Эволюционные базы данных и знаний для адаптивного синтеза интеллектуальных систем. Миварное информационное пространство. - М.: Радио и связь, 2002. – 288 с.

[5] Варламов О.О. Обзор 25 лет развития миварного подхода к разработке интеллектуальных систем и создания искусственного интеллекта // Труды НИИР. 2011. № 1. С. 34-44.

[6] Варламов О.О., Сергушин Г.С., Елисеев Д.В., Адамова Л.Е., Майборода Ю.И., Антонов П.Д., Чибирова М.О. О миварном подходе к моделированию процессов понимания компьютерами смысла текстов, речи и образов. Новые возможности расширения границ автоматизации умственной деятельности человека. // Автоматизация и управление в технических системах. – 2013. – № 2. – С. 30-45.

[7] Варламов О.О., Адамова Л.Е., Петерсон А.О., Протопопова Д.А., Скакунова Е.А. Исследование подходов и основных проблем понимания естественного русского языка // Автоматизация и управление в технических системах. – 2014. – № 2; URL: auts.esrae.ru/10-196 (дата обращения: 15.09.2014).

[8] Варламов О.О. Системный анализ и синтез моделей данных и методы обработки информации в самоорганизующихся комплексах оперативной диагностики: диссертация на соискание ученой степени доктора технических наук. – М.: МАРТИТ, 2003. 307 с.

[9] Варламов О.О. Разработка адаптивного механизма логического вывода на эволюционной интерактивной сети гиперправил с мультиактивизаторами, управляемой потоком данных // Искусственный интеллект. 2002. № 3. С. 363-370.
[10] Варламов О.О. Основы многомерного информационного развивающегося (миварного) пространства представления данных и правил // Информационные технологии, 2003. № 5. С. 42-47.
[11] Варламов О.О. Разработка метода распараллеливания потокового множественного доступа к общей базе данных в условиях недопущения взаимного искажения данных // Информационные технологии. 2003. №1. С. 20-28.
[12] Варламов О.О. Параллельная обработка потоков информации на основе виртуальных потоковых баз данных // Известия высших учебных заведений. Электроника. 2003. № 5. С. 82-89.
[13] Варламов О.О. Системный анализ и синтез моделей данных и методы обработки информации для создания самоорганизующихся комплексов оперативной диагностики // Искусственный интеллект. 2003. № 3. С. 299-305.
[14] Варламов О.О. Системы обработки информации и взаимодействие групп мобильных роботов на основе миварного информационного пространства // Искусственный интеллект. 2004. № 4. С. 695-700.
[15] Варламов О.О. Создание интеллектуальных систем на основе взаимодействия миварного информационного пространства и сервисно-ориентированной архитектуры // Искусственный интеллект. 2005. № 3. С. 13-17.
[16] Варламов О.О. Анализ взаимосвязей GRID и САС ИВК, SOA и миварного подхода // Искусственный интеллект. 2005. № 4. С. 4-11.
[17] Максимова А.Ю., Варламов О.О. Миварная экспертная система для распознавания образов на основе нечеткой классификации и моделирования различных предметных областей с автоматизированным расширением контекста // Известия Южного федерального университета. Технические науки. 2011. № 12. С. 77-87.
[18] Варламов О.О. О необходимости перехода от теории искусственного интеллекта к разработке теории активного отражения // Известия Южного федерального университета. Технические науки. 2007. Т. 77. № 2. С. 89-95.
[19] Владимиров А.Н., Варламов О.О., Носов А.В., Потапова Т.С. Программный комплекс "УДАВ": практическая реализация активного обучаемого логического вывода с линейной вычислительной сложностью на основе миварной сети правил // Труды НИИР. 2010. Т. 1. С. 108-116.
[20] Варламов О.О. Миварные технологии: переход от продукций к двудольным миварным сетям и практическая реализация автоматического конструктора алгоритмов, управляемого потоком входных данных и обрабатывающего более трех миллионов продукционных правил // Искусственный интеллект. 2012. № 4. С. 11-33.
[21] Варламов О.О. Практическая реализация линейной вычислительной сложности логического вывода на правилах "ЕСЛИ-ТО" в миварных сетях и обработка более трех миллионов правил // Автоматизация и управление в технических системах. – 2013. – № 1; [Электронный ресурс]. URL: http://auts.esrae.ru/3-66 (дата обращения: 26.03.2013).
[22] Варламов О.О., Чибирова М.О., Сергушин Г.С., Елисеев Д.В. "Облачная" реализация миварного универсального решателя задач на основе адаптивного активного логического вывода с линейной сложностью относительно правил "Если-То-Иначе" // Автоматизация и управление в технических системах. – 2013. – № 2. С. 7-23.

[23] Сергушин Г.С., Варламов О.О., Чибирова М.О., Елисеев Д.В., Муравьева Е.А. Исследование возможностей информационного моделирования сложных систем управления технологическими процессами на основе миварных технологий // Автоматизация и управление в технических системах. – 2013. – № 2. С. 46-60.

[24] Варламов О.О., Адамова Л.Е., Елисеев Д.В., Майборода Ю.И., Антонов П.Д., Сергушин Г.С., Чибирова М.О. Комплексное моделирование процессов понимания компьютерами смысла текстов, речи и образов на основе миварных технологий // Искусственный интеллект. – 2013. – № 4. – С. 15-27.

[25] Чибирова М.О., Сергушин Г.С., Варламов О.О., Елисеев Д.В., Хадиев А.М. и др. Реализация общедоступного миварного универсального решателя задач на основе адаптивного активного логического вывода с линейной сложностью и облачных технологий // Искусственный интеллект. – 2013. – № 3. – С. 512-523.

[26] Белоусова А.И., Варламов О.О., Остроух А.В., Краснянский М.Н. Подход к формированию многоуровневой модели мультиагентной системы с использованием миваров // Перспективы науки. 2011. № 20. С. 57-61.

[27] Варламов О.О., Владимиров А.Н., Бадалов А.Ю., Чванин О.Н. Развитие миварного метода логико-вычислительной обработки информации для АСУ, тренажеров, экспертных систем реального времени и архитектур, ориентированных на сервисы // Труды Научно-исследовательского института радио. 2010. № 3. С. 18-26.

[28] Владимиров А.Н., Варламов О.О., Носов А.В., Потапова Т.С. Применение многопроцессорного вычислительного кластера НИИР для распараллеливания алгоритомв в научно-технических и вычислительных задачах // Труды Научно-исследовательского института радио. 2009. № 3. С. 120-123.

[29] Варламов О.О. Миварный подход к разработке интеллектуальных систем и проект создания мультипредметной активной миварной интернет-энциклопедии // Известия Кабардино-Балкарского научного центра РАН. 2011. № 1. С. 55-64.

[30] Подкосова Я.Г., Варламов О.О., Остроух А.В., Краснянский М.Н. Анализ перспектив использования технологий виртуальной реальности в дистанционном обучении // Вопросы современной науки и практики. Университет им. В.И. Вернадского. 2011. № 2. С. 104-111.

[31] Варламов О.О. Эволюционные базы данных и знаний. Миварное информационное пространство // Известия Южного федерального университета. Технические науки. 2007. Т. 77. № 2. С. 77-81.