АвиаОнтология: анализ современного состояния ресурса[1]

Н.В. Лукашевич
НИВЦ МГУ; АНО Центр информационных исследований,
louk@mail.cir.ru

О.А. Невзорова
НИИММ им. Н.Г. Чеботарева; Казанский государственный педагогический университет,
Olga.Nevzorova@ksu.ru

Статья посвящена анализу современного состояния нового информационного ресурса АвиаОнтологии. Онтология была разработана на основе технологии создания больших и сверхбольших онтологий и тезаурусов для различных областей в НИВЦ МГУ. В настоящее время АвиаОнтология включает около 1600 понятий и 4700 терминов. В статье мы рассмотрим структурные характеристики онтологии и визуальную модель ее представления, выполненной с помощью системы “OntoEditor”.

1. Введение

Информационный ресурс АвиаОнтология является лингвистической онтологией, разработанной для специальной прикладной области. Данная прикладная область включает знания об авиации и специализирована на знаниях об информационных событиях и процессах в авиационной практике. Прежде всего это касается событий и процессов функционирования бортовой аппаратуры, а также экипажа в различных полетных режимах.

Необходимость разработки онтологии была вызвана прикладными задачами автоматической обработки специализированных текстов данной предметной области. Одним из приложений АвиаОнтологии является система автоматического анализа специальных технических текстов типа «Логика работы…» [1]. Задачи приложения определили специализацию АвиаОнтологии, которая связана, в первую очередь, с расширенным описанием области процессов обмена и передачи информации при решении задач авиационной тематики.

К настоящему времени АвиаОнтология содержит 1600 понятий, 4700 текстовых единиц. Проектирование подобной модели является сложной экспертной задачей, требует взвешенного подхода к выбору формализма описания.

[1] Данное исследование выполнено при поддержке Российского Фонда Фундаментальных исследований, грант № 02-07-90279.

В качестве базовой технологии для проектирования прикладной онтологии была выбрана технология проектирования больших и сверхбольших онтологий и тезаурусов для различных предметных областей [2], разработанная в рамках проекта Университетская информационная система РОССИЯ (www.cir.ru), поддерживаемая Научно-исследовательским вычислительным центром МГУ им. М.В.Ломоносова и АНО Центр информационных исследований.

2. Структура АвиаОнтологии

Разработка АвиаОнтологии осуществляется на основе электронной коллекции текстов данной предметной области, которая в настоящий момент составляет свыше 100 Мб. Основными текстовыми источниками знаний о рассматриваемой предметной области являются серия специальных книг, отобранных экспертами в данной области, и текстовая информация из Интернет по отдельным разделам предметных знаний.
АвиаОнтология представляет иерархическую сеть понятий, и ее проектирование осуществлялось в несколько этапов.
На первом этапе осуществлялось построение "терминологического портрета" предметной области на основе автоматических методик. Специализированные алгоритмы автоматического выделения терминологических словосочетаний, в том числе и многословных терминов, в текстах электронной коллекции описаны в [3]. Использование различных фильтров при обработке списка кандидатов в термины (около 14 000 слов и словосочетаний) позволило автоматически выделить около 700 основных терминов предметной области.
Терминология любой предметной области содержит как специфические термины, употребляемые только в данной области или в ряде близких областей, так и достаточно общеизвестные термины. В данной области таким общеизвестными терминами являются *летчик, самолет, истребитель, оружие, боевые действия, атака* и многие другие. Это позволяет не начинать разработку модели предметной области с нуля, а использовать знания, описанные в более общих лингвистических и терминологических ресурсах. В качестве такого ресурса был использован тезаурус РуТез, разработанный АНО Центр информационных исследований.
Тезаурус представляет собой иерархическую сеть понятий, каждое из которых имеет ряд текстовых вариантов (способов языкового выражения) и совокупность отношений с другими понятиями тезауруса. Объем ресурса на момент проектирования онтологии составлял 97 тысяч слов и словосочетаний, уложенных в 42 тысячи понятий. Между понятиями вручную установлено более 160 тысяч связей. По свойствам транзитивности и наследования выводится более 1200000 связей между понятиями.
Наличие большого общезначимого лингвистического ресурса позволило сопоставить данный ресурс с текстами в предметной области, выделить описанные в тезаурусе типы знаний (понятия, синонимы, связи между понятиями), перенести их в специальную рабочую область как основу для построения прикладной онтологии.
Списки набранных слов и словосочетаний предметной области были сопоставлены с терминами Общественно-политического тезауруса. Если сопоставление очередного словосочетания было успешно, соответствующее понятие Общественно-политического тезауруса вместе со всеми терминами, выражающими его в тексте, копировалось в модель предметной области. На следующем шаге были скопированы все отношения Общественно-политического тезауруса между скопированными понятиями.
Кроме того, было выполнено замыкание отношений - выбирались не только те понятия, которые упомянуты в текстах предметной области непосредственно, но и те понятия

тезауруса, которые находятся на концептуальных путях между упомянутыми в текстах понятиями - если понятие В является вышестоящим для понятия А, понятие С является вышестоящим для понятия В, причем понятия А и С были скопированы в предметную область, то и понятие В также копируется в модель предметной области вместе со своими терминами-вариантами и релевантными отношениями.

Безусловно, перенесенные понятия и отношения требуют тщательного дополнительного тестирования для задания настройки на конкретную область. Например, в результате переноса среди терминов может встретиться синонимический вариант, который маловероятен в данной области, например, словосочетание *винтокрылая машина* как синоним слова *вертолет* вряд ли может встретиться в профессиональной технической области. Поэтому был осуществлен ручной контроль перенесенных терминов.

В существующей версии Авиа-Онтологии около 500 понятий, 1000 синонимов и 1000 отношений было перенесено из тезауруса РуТез, что обеспечило быстрый рост Авиа-Онтологии и существенно сократило время разработки нового ресурса.

Тестирование АвиаОнтологии на специальных текстах, при которых анализировалось покрытие текстов терминами из онтологии, показало, что для задач прикладного использования АвиаОнтологии требуется ввести в онтологию важные общезначимые слова, такие как *необходимость, вероятность, условие* и т.п. Группа из 130 таких понятий была добавлена в АвиаОнтологию.

Следующий этап проектирования онтологии - проектирование сети отношений онтологических концептов. Ключевой проблемой данного этапа является выбор списка отношений между концептами. В качестве базовой системы отношений была использована система отношений тезауруса РуТез, которая включает иерархическое отношение (ВЫШЕ-НИЖЕ), отношение ЧАСТЬ-ЦЕЛОЕ, отношения онтологической зависимости (АСЦ, АСЦ1, АСЦ2).

Отношения ЧАСТЬ – ЦЕЛОЕ – используется для описания как традиционных частей, так и участников ситуаций, свойств.

Например,

Самолет заправщик
Целое ДОЗАПРАВКА В ВОЗДУХЕ
ФЮЗЕЛЯЖ
Целое САМОЛЕТ
ЛЕТНО-ТЕХНИЧЕСКИЕ ХАРАКТЕРИСТИКИ
Целое ЛЕТАТЕЛЬНЫЙ АППАРАТ

Характерной онтологической особенностью всех этих достаточно традиционных видов отношений является то, что ЦЕЛОЕ (носитель свойства, ситуация) обычно онтологически зависит [4] от своих ЧАСТЕЙ (свойств, участников), имеет таких ЧАСТЕЙ (свойств, участников) более одного, и эти ЧАСТИ (свойства, участники) могут онтологически не зависеть друг от друга.

Для обеспечение транзитивности логического вывода по отношению ЧАСТЬ – ЦЕЛОЕ требуется, чтобы ЧАСТИ также онтологически зависели от ЦЕЛОГО. Чаще всего такое описание возможно. Например, не следует в АвиаОнтологии считать, что *ДВИГАТЕЛЬ* это ЧАСТЬ *САМОЛЕТА*, двигатели бывают также, например, в ракетах. Это в данном случае и означает, что существование понятия *ДВИГАТЕЛЬ* не зависит от существования понятия *САМОЛЕТ*. Таким образом, необходимо выделить вид двигателя *АВИАЦИОННЫЙ ДВИГАТЕЛЬ*, который и будет описан как ЧАСТЬ понятия *САМОЛЕТ*.

Для описания других отношений онтологической зависимости используется несимметричная ассоциация АСЦ1-АСЦ2 (АСЦ1="зависит_от", АСЦ2="главное_для");

Симметричная ассоциация используется для сходных по смыслу понятий.

Приведем пример словарной статьи Авиа-Онтологии, содержащей все типы отношений, с пометкой 'син' даны синонимы понятия:

ПУСК РАКЕТЫ

син	*запуск ракеты*
син	*запустить ракету*
син	*отстрел ракеты*
син	*применение ракет*
син	*применение ракетного вооружения*
син	*пуск ракеты*
син	*пустить ракету*
син	*ракетная атака*
син	*ракетный удар*
ВЫШЕ	*ПРИМЕНЕНИЕ ОРУЖИЯ*
НИЖЕ	*ЭФФЕКТИВНЫЙ ПУСК РАКЕТЫ*
ЧАСТЬ	*ДАЛЬНОСТЬ ПУСКА РАКЕТЫ*
ЧАСТЬ	*ЗОНА ВОЗМОЖНОГО ПУСКА*
ЧАСТЬ	*МАКСИМАЛЬНО ДОПУСТИМАЯ ПЕРЕГРУЗКА ПРИ ПУСКЕ*
АСЦ1	*РАКЕТА*
АСЦ2	*ВЫХОД В ЗОНУ ВОЗМОЖНОГО ПУСКА*
АСЦ2	*КОМАНДА «ПУСК РАЗРЕШЕН»*
АСЦ2	*ОШИБКА ПУСКА*
АСЦ2	*УПРАВЛЕНИЕ РАКЕТОЙ*

3. Визуализация АвиаОнтологии

Исследование характеристик АвиаОнтологии было выполнено с помощью инструментальной системы визуального проектирования онтологий «OntoEditor» разрабатываемой в НИИ математики и механики им. Н.Г. Чеботарева г. Казань, руководитель проекта О.А. Невзорова. Проектирование информационных ресурсов типа онтологий на основе специализированной программной системы «OntoEditor» осуществляется на основе визуальной технологии, позволяющей эксперту или инженеру по знаниям вводить понятия, синонимические ряды понятий, связи между понятиями и отображать введенную информацию в графическом режиме.

Инструментальная система визуального проектирования «OntoEditor» является специализированной СУБД. Система предназначена для ручного редактирования онтологий, хранящихся в реляционной базе данных в определенном формате, а также обслуживания запросов пользователей и внешних программ.

Инструментальная система позволяет:

- добавлять, изменять и удалять отдельные записи БД;
- автоматически корректировать данные при удалении конкретных записей (например, удалять отношения стертого концепта);
- поддерживать ведение нескольких онтологий, в том числе смешанных (например, с общими списками типов отношений, классов, синонимов и др.);
- импортировать онтологии различных форматов данные из внешних баз данных (механизмы импорта разрабатываются на конкретную базу данных);
- вести обработку онтологий в табличной и графической формах;
- поддерживать иерархическую структуру классов;
- выделять по заданному фильтру определенные подмножества редактируемой онтологии;
- вести автоматическую статистику по объектам онтологии;
- осуществлять поиск цепочек отношений концептов с заданными свойствами;

- выводить требуемую информацию на печать или в текстовый файл в заданных формах отчета в кодировках ANSI и ASCII;
- обрабатывать внешние запросы с использованием механизма обмена DDE (Dynamic Data Exchange).

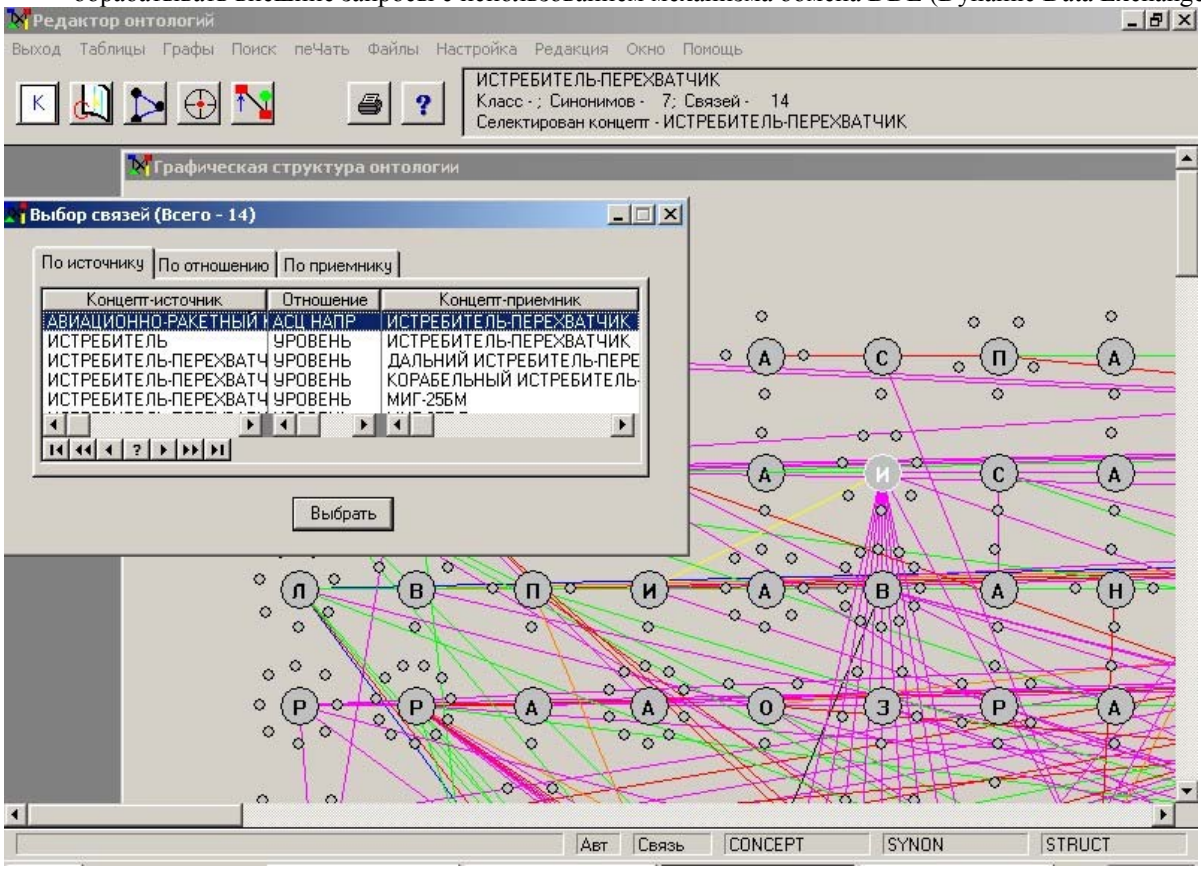

Рис.1. Визуальный образ АвиаОнтологии

Система поддерживает многооконный интерфейс и снабжена развитой системой подсказок, а также механизмами поиска конкретной записи.

На рис. 1 показан визуальный образ АвиаОнтологии с селектированным концептом *ИСТРЕБИТЕЛЬ* (выделен цветом на форме) и окном отображения связей селектированного концепта. Синонимы концепта размещаются на окружности с центром-концептом, типы отношений выделяются цветом.

С помощью инструментальной системы визуального проектирования «OntoEditor» были получены графические образы АвиаОнтологии с различными структурными характеристиками. Ранжирование АвиаОнтологии по различным критериям (по числу синонимов, по количеству связей и др.) позволили получить различные “графические портреты”. Так, при ранжировании по синонимам получено распределение концептов по количественным уровням синонимов (Рис.2). Синоним концепта определяет отношение “концепт - текстовый вход концепта”. Синонимы задают текстовые входы концепта, не различимые в ситуациях их использования в текстах. Анализ гистограммы распределения концептов показывает, что максимальным числом синонимом обладают концепты *СООБЩИТЬ (УВЕДОМИТЬ)*, *ПРИМЕНЕНИЕ ОРУЖИЯ*, *ВЫПОЛНИТЬ (ИСПОЛНИТЬ, ОСУЩЕСТВИТЬ)*.

Рис.2. Гистограмма распределения синонимов концептов

Рис.3. Гистограмма распределения связей концептов

Пример синонимического ряда концепта *ПРИМЕНЕНИЕ ОРУЖИЯ*:
(*боевое применение; ведение огня; вести огонь; нанесение удара; нанести удар; обстрел; обстреливать; обстреливаться; обстрелять; огневое воздействие; огонь; открывать огонь; открыть огонь; применение средств вооружений; применить оружие; простреливать; стрельба; стрелять; стрельнуть;*).
При ранжировании по количеству связей получено распределение концептов по количественным уровням связей (рис.3). Наибольшее количество связей (рассматривается степень вершины) имеет концепт *ИСТРЕБИТЕЛЬ* – 30 связей. При этом состав установленных связей следующий: 23 иерархических отношения, 6 – отношений направленной ассоциации и одно отношение ЧАСТЬ-ЦЕЛОЕ. Для концепта *САМОЛЕТ* установлено 23 связи, из которых 10 иерархических отношений, 5 – отношений ЧАСТЬ-ЦЕЛОЕ, 8 – отношений направленной ассоциации.
Полученный результат отражает общие принципы установления отношений, существенное использование механизма наследования по транзитивным связям.

Концепты *САМОЛЕТ* и *ИСТРЕБИТЕЛЬ* связаны минимальной цепочкой иерархических отношений *САМОЛЕТ - РЕАКТИВНЫЙ САМОЛЕТ -ИСТРЕБИТЕЛЬ*, что позволяет концепту *ИСТРЕБИТЕЛЬ* наследовать свойства от концепта *САМОЛЕТ*, и в свою очередь детализировать собственное описание за счет существенного введения видовых объектов (типов истребителей) по отношению НИЖЕ.

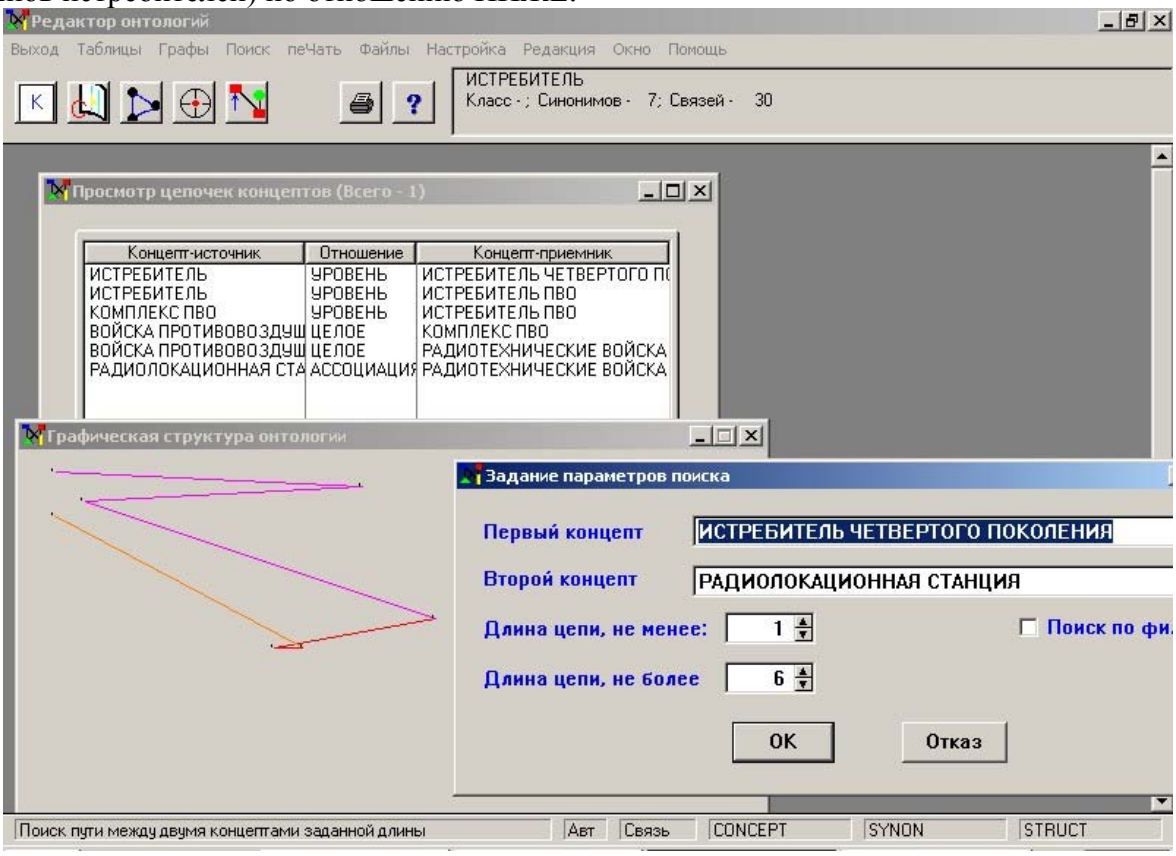

Рис.4. Поиск пути между концептами заданной длины

Особый интерес представляет анализ цепочек отношений концептов АвиаОнтологии. При построении цепочки учитывается направление связи, т.е. допустимая цепочка концептов является путем (полупуть не разрешается). Механизм построения цепочек поддерживает вывод на онтологии, который используется в прикладных задачах. При этом исследуется вопрос о структуре вывода, т.е. о числе шагов вывода, промежуточные концептах в цепочке вывода, типах связей. Очевидно, что достоверный вывод обеспечивает вывод по транзитивным связям, где не важно количество промежуточных звеньев. Структуры вывода с ассоциативными связями и вывод с перегибами по транзитивным связям требуют дополнительного изучения.

На рис. 4 приведена структура пути между концептами *ИСТРЕБИТЕЛЬ ЧЕТВЕРТОГО ПОКОЛЕНИЯ* и *РАДИОЛОКАЦИОННАЯ СТАНЦИЯ*. Минимальная длина пути – 6 связей. На форме также приведен графическая структура вывода при ранжировании концептов по номерам в базе.

Заключение

Текущая версия АвиаОнтологии представляет собой специализированный информационный ресурс, который уже на данном состоянии разработки может быть встроен в различные приложения. К числу потенциальных приложений можно отнести ряд информационно-поисковых задач, таких как содержательный поиск документов с

расширением запроса по иерархии онтологии, нахождения похожих документов, классификации документов по одному или нескольким классификаторам, например, классификации по ситуациям Обороны, Атаки, Посадки, Взлета и т.п. Другие приложения связаны с задачами анализа специализированных текстов.
Развитие АвиаОнтологии предусматривает ее дальнейшую специализацию. Отдельным направлением исследований является расширение типов отношений онтологии, добавление отношений последовательностей (ситуационных и временных).
Визуальные методы проектирования онтологий способствуют более быстрому и более полному пониманию структуры знаний предметной области, что особенно ценно для исследователей, работающих в новой предметной области.
Инструментальная система «OntoEditor» предоставляет эффективный набор инструментальных средств, позволяющих осуществлять проектирование онтологии любой структуры, с любыми типами отношений и любыми классами концептов. Особенно важно, что система поддерживает импорт любого типа онтологии, реализованной на физическом уровне в табличной форме. Поддержка разнообразных поисковых запросов и механизмы вывода на онтологии позволяют исследовать внутреннюю структуру знаний предметной области. Система «OntoEditor» может быть встроена в приложения различного назначения, работающие с большими базами знаний.

Литература

1. Невзорова О.А., Федунов Б.Е., Система анализа технических текстов "ЛоТА": основные концепции и проектные решения. // Изв. РАН. Теория и системы управления. – 2001. – № 3.– С. 138-149.
2. Лукашевич Н.В., Добров Б.В., Тезаурус русского языка для автоматической обработки больших текстовых коллекций // Компьютерная лингвистика и интеллектуальные технологии: Труды Международного семинара Диалог’2002 / Под ред. А.С.Нариньяни – М.: Наука – 2002. – Т.2 - С.338-346.
3. Dobrov B., Loukachevitch N., Nevzorova O. An approach to new ontologies development: main ideas and simulation results //Int. Journal Information Theories & Applications. Vol.10. Number 1, 2003. P.98-105.
4. Guarino N., Some Ontological Principles for Designing Upper Level Lexical Resources // Proceedings of First International Conference on Language Resources and Evaluation, 1998..