сертификатов за определенный период времени. При превышении заданного лимита ЦС отклоняет запрос на выдачу ЭЦП.

Администратору необходимо настроить логику работы почтового сервера при превышении лимита ЦС. Если лимит исчерпан, письмо либо отправляется без ЭЦП, либо его пересылка прекращается или откладывается до получения ЭЦП. Во всех случаях администратор получает сообщение о превышении заданного лимита на отправление электронных писем, что позволяет оперативно выявить причину случившегося.

Построенные модели показали, что использование ЭЦП при передаче письма увеличивает на 25% загрузку канала связи для письма с минимальной длиной. С увеличением объема письма этот процент снижается. Вместе с тем значительно снизится объем передаваемой корреспонденции за счет фильтрации спама.

4. Заключение

Использования более действенных механизмов фильтрации корреспонденции на основе ЭЦП предотвратит несанкционированные массовые рассылки.

5. Литература

RFC 2821, Klensin, J.,
Simple Mail Transfer Protocol, April, 2001

M. Sahami, S. Dumais, D. Heckerman, E. Horvitz,
A Bayesian approach to 'ltering junk E-mail, in: AAAI Workshop on Learning for Text Categorization Madison, Wisconsin. AAAI Technical Report WS-98-05, July, 1998

УДК 681.3.06

МЕТОДЫ, МАТЕМАТИЧЕСКАЯ МОДЕЛЬ ОПИСАНИЯ И ПРОГРАММНЫЕ СРЕДСТВА АВТОМАТИЗАЦИИ ПОСТРОЕНИЯ ЛЕКСИКО-СЕМАНТИЧЕСКОЙ БАЗЫ ДАННЫХ WORDNET

А.М. Сухоногов

Аннотация

Для решения ряда задач, таких как информационный поиск (information retrieval), машинный перевод (machine translation), определение значения слов (WSD – word-sense-disambiguation), при построении вопросно-ответных систем (Q&A systems) в качестве базы знаний, может использоваться лексико-семантический словарь WordNet. Разработка WordNet для русского языка является актуальной задачей. В работе рассматриваются особенности реализации русского WordNet, описывается методика его построения и проверки. Предложенная методика позволяет значительно сократить время разработки словаря за счет более эффективного использования доступных лексических ресурсов, позволяет контролировать каждый этап его построения. Предложенный формат представления WordNet в виде лексической онтологии на языке OWL (Ontology Web Language) позволяет интегрировать словарь с другими словарями, а также использовать его в качестве компонента технологии SemanticWeb.

Ключевые слова: WordNet; лексико-семантическая база данных; тезаурус; онтология; Semantic Web; RDF; OWL

Введение

Лексико-семантические словари WordNet получили широкое распространение после появления в 1996 году в свободном доступе Принстонского WordNet английского языка (Fellbaum, 1998). В 1999 году был создан многоязычный вариант WordNet – EuroWordNet, объединивший национальные словари WordNet европейских языков (английский, датский, испанский, итальянский, немецкий, французский, чешский и эстонский). В 2004 году завершается работа над проектом BalkaNet, объединяющем греческий, болгарский, турецкий, чешский, французский, румынский и сербский языки. В 2001 году создана всемирная ассоциация – Global WordNet Association, координирующая разработки национальных WordNet словарей.

Близкие задачи решались при построении теории лингвистических моделей типа Смысл-Текст. Но определенная там концепция толково-комбинаторного словаря (Мельчук И.А, 1974) не получила широкого распространения.

Разработка русского WordNet осуществляется коллективом из сотрудников ПГУПС (каф.ИВС) и компании Руссикон под руководством Яблонского С.А. Основная цель проекта – разработка русско-английского словаря WordNet и обеспечение его интеграции с другими словарями.

1. Структура WordNet

Базовой структурной единицей WordNet является синонимичный ряд – синсет, объединяющий слова со схожим значением. Каждый синсет представляет некоторое лексикализованное понятие языка. Для удобства пользования словаря человеком каждый синсет дополнен определением и примерами употребления слов в контексте. Синсеты в WordNet связаны между собой семантическими отношениями гипонимии (род/вид), меронимии (часть/целое), лексического вывода (каузации, пресуппозиции)

и др. Определяются также отношения и между значениями слов, например антонимия.

Для русского WordNet разработана структура, позволяющая дополнительно определять парадигму каждого слова, стилистические пометы, пометы доминантности. По результатам анализа проектов EuroWordNet (Vossen, 1999) и BalkaNet (Horák и др., 2003) разработан универсальный формат представления WordNet на языке определения онтологий – OWL. OWL-описание WordNet доступно по запросу (ASukhonogov@rambler.ru).

2. Разработка русского WordNet

2.1. Ресурсы для построения русского WordNet

Для построения русского WordNet используются лингвистические ресурсы компании Руссикон и словари, свободно распространяемые в Internet. Научный коллектив из сотрудников ПГУПС (каф.ИВС) и компании Руссикон выиграл в 2003 г. конкурс издательства Oxford Press на лучший исследовательский проект по использованию словарей Oxford Press. В настоящее время издательство Oxford Press представило для создания русской версии WordNet XML версии следующих словарей:

- Oxford Russian Dictionary
- New Oxford Dictionary of English, 2nd Edition
- New Oxford Thesaurus of English

Эти ресурсы будут использоваться при автоматизированном построении межъязыкового индекса (ILI-inter-lingual-index) русско-английского WordNet.

2.2. Основные этапы разработки

Разработка русско-английского WordNet состоит из двух этапов. На первом этапе формируется WordNet словарь русского языка, второй этап заключается в построении межъязыкового индекса и русско-английского WordNet.

Для разработки русского WordNet определены следующие принципы:

- использование всей доступной в лексических ресурсах информации для автоматизации процесса построения словаря;
- полный контроль и протоколирование всех преобразований на каждом этапе. Прозрачность преобразований;
- минимизация ручной обработки результатов каждого этапа или исключение такой обработки;
- автоматическое отслеживание и исключение несоответствий на этапах сборки и редактирования WordNet.

Традиционный подход к построению WordNet предполагает анализ большого корпуса текстов, построение частотных словарей, их анализ, выявление наиболее употребительной лексики. Ядро WordNet включает

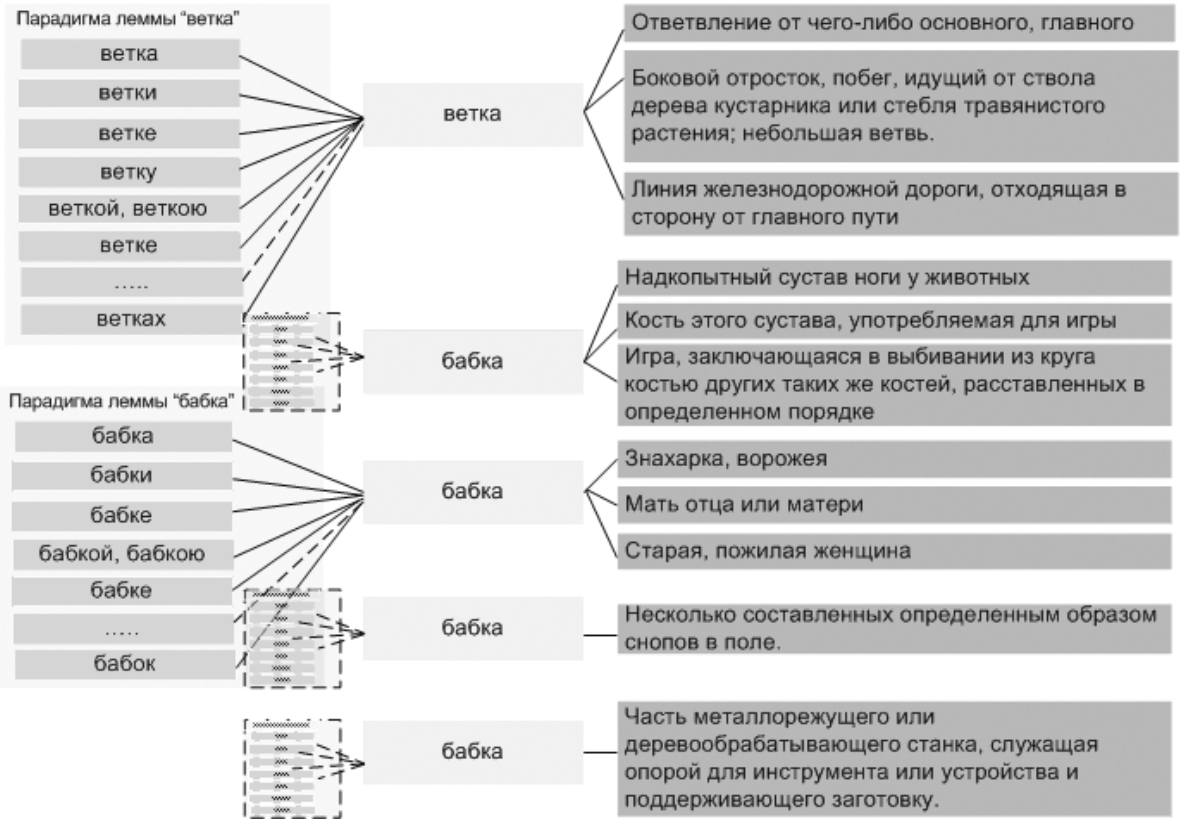
наиболее общие концепты и системы их отношений, для каждого синсета определяются дефиниций. Далее приступают к обработке менее употребляемой лексики и определению предметных областей.

Наличие готовых словарей позволяет оптимизировать процесс построения WordNet. При таком подходе появляются задачи определения соответствия между статьями словарей-источников. Такое соответствие не всегда однозначно.

2.3. Основные задачи и алгоритмы

2.3.1. Анализ толкового словаря

Алгоритм анализа толкового словаря позволяет выделить значения слов и другую вспомогательную информацию (стилистические пометы, примеры использования, некоторые грамматические характеристики, отношения словообразования, позиции ударных гласных и т.д.) из статей толкового словаря. На рис.1. приводиться пример прототипов статей синсетов русского WordNet, формируемых при обработке лемм «ветка» и



«бабка» толкового словаря.

Рис.1. Фрагмент толкового словаря

2.3.2. Объединение толкового и грамматического словарей

Алгоритм позволяет автоматизировать процесс определения грамматических характеристик полученных из толкового словаря лемм.

Дальнейшая обработка журналов регистрации объединений позволяет дополнить существующий грамматический словарь и исключить из дальнейшей обработки леммы, попавшие в словник по ошибке. Результаты работы такого алгоритма представлены на рис. 2. По результатам обработки видно, что фактически проверке подлежит около 15% всего словника. А сама проверка заключается в расстановке простых помет определяющих операции по исправлению словаря.

2.3.3. Формирование синсетов по пометам толкового словаря

Алгоритм позволяет сформировать полноценные синсеты WordNet, если исходные лексические ресурсы содержат данные о наличии синонимичных отношений. При определении синсетов учитывается наличие нескольких значений для каждой из лемм. Проводятся дополнительные проверки. Например, проверяется соответствие частей речи объединяемых лемм. Отслеживаются цепочки, включающие последовательное определение нескольких синонимов. Пример одного из вариантов такого объединения показан на рис.3.

Рис.2. Результат работы алгоритма объединения словарей

2.3.4. Объединение прототипа WordNet со словарем синонимов

Результатом работы является набор синсетов, доопределенных синонимами и система родовидовых отношений между ними. При определении соответствия учитываются практически все доступные компоненты словарных статей: словарный состав статей, толкования, примеры использования (речения и фразеологизмы) и т.д.

Рис.3. Удачное разрешение задачи формирования синсета

3. Заключение

В настоящее время русский WordNet объединяет более 100 тыс. синсетов и находиться на этапе проверки. Для построения, проверки и редактирования разработан набор утилит и редактор “TenDrow”. Разработаны процедуры экспорта/импорта форматы редакторов Princeton WordNet и BalkaNet (VisDic-XML). Разработан собственный формат представления на языке OWL. Данные проекта доступны в Internet.

Для получения универсального алгоритма быстрого формирования национального WordNet, достаточно договорится о единой форме представления входных и выходных данных для каждого алгоритма. В данной реализации таким форматом был формат представления словарей «Руссикон». Наиболее удачной реализацией является формат представления XML. Недостатком методики является необходимость наличия лексических ресурсов - словарей в электронном виде и в заранее определенном формате.

4.Литература

Christiane Fellbaum (ed.), WordNet: An Electronic Lexical Database, MIT Press, 1998

Мельчук И.А. Опыт теории лингвистических моделей «Смысл-Текст». Семантика, синтаксис. Москва: Наука. 1974

Piek Vossen. EuroWordNet. General Document. Version 3. University of Amsterdam, http://www.hum.uva.nl/~ewn

Aleš Horák and Pavel Smrž “VisDic – Wordnet Browsing and Editing Tool”. Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings, pp. 136–141. Masaryk University, Brno, 2003