Многоязычная лингвистическая база знаний: архитектура и метаданные

© Н.В. Лунева

Институт проблем информатики
nl2@mail.ru

Аннотация

Данная работа содержит описание основных архитектурных решений и типов метаданных в многоязычной лингвистической базе знаний, создаваемой для построения и отладки синтактико-семантических моделей в лингвистических процессорах систем обработки текстовых знаний и машинного перевода. Новая база знаний предназначена для широкого круга специалистов в области компьютерной лингвистики и информационных технологий. В рамках комплекса предусмотрено создание рабочего места переводчика и компоненты «переводческой памяти».

1 Введение

Основная часть знаний, которыми сегодня располагает человечество, представлена в виде текстов на естественных языках. В связи с этим построение многоязычной лингвистической базы знаний, дополненной функциями машинного перевода, направленной на формирование единого семантического представления полного текста научных документов на различных языках, служит инструментом повышения качества и эффективности решения информационных задач.

Многоязычная лингвистическая база знаний создается на основе параллельных научных и патентных текстов и метаязыковых лингвистических представлений для систем машинного перевода и обработки текстовых знаний [9]. Исследования и разработка проводятся на базе внутреннего лингвистического ресурса «Полиглот» и направлены на создание инженерно-лингвистической среды, включающей в себя лингвистическую базу знаний и функции разбора и перевода языковых структур.

Работа над многоязычной лингвистической базой знаний проводится в русле наиболее актуальных мировых тенденций по созданию семантически-ориентированных лингвистических ресурсов и систем инструментальной и информационной поддержки решения инженерно-лингвистических задач и автоматизации перевода, развиваемых отечественными и зарубежными исследователями и разработчиками [1,3,6,7,8,13]. В частности, нами были рассмотрены лингвистические базы данных, исследованы функциональные характеристики и архитектурные решения систем WordNet, EuroWordNet, RussNet, проводились сопоставления с системой Alchemy Catalyst, в которой предусмотрен набор инструментов для организации и информационного обеспечения переводческой деятельности.

2 Многоязычная лингвистическая база знаний

Многоязычная лингвистическая база знаний создается на основе экспериментальной базы данных для отладки семантико-синтаксических представлений в лингвистических процессорах систем машинного перевода и обработки текстовых знаний. При разработке лингвистического процессора (на основе англо-русского и обратного трансфера) Е.Б. Козеренко [9] было предложено понятие полей функционального переноса, ставших базисом сегментации языковых структур для решения задач машинного перевода. Основная идея такого поля состоит в принятии гипотезы о том, что в основе грамматических структур лежат когнитивные структуры (ментальные фреймы); функционально-семантическое поле отражает взаимодействие элементов разных языковых уровней [5,9].

В основе многоязычной лингвистической базы знаний лежит разработка системы правил фразовых структур, отражающих также и отношения зависимости через механизм наследования атрибутов головной вершины [12]. Как показано в [9], этот подход более практичен с вычислительной точки зрения, и не применялся ранее для двуязычной ситуации. Функциональные значения языковых единиц закодированы как метки фразовых структур, и типы атрибутов-значений определяются функционально-категориальной семантикой [5,9]. Множество языковых структур, представленных в виде синтактико-семантических комплексов,

Труды 9ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2007, Переславль-Залесский, Россия, 2007.

выстраиваются в иерархию правил. Отношения зависимости реализуются через механизм головных вершин фразовых структур, а сами фразовые структуры задают линейные последовательности языковых объектов [4,10].

На текущем этапе система правил модифицируется с учетом возможной многозначности синтаксических структур, и разрабатываются механизмы разрешения неоднозначности посредством включения в систему правил статистической информации о возможных контекстах языковых структур. Планируется включение механизмов обучения для расширения и модификации правил. Для того, чтобы избежать порождения избыточных правил, в систему закладываются не только шаблоны языковых структур, по аналогии с которыми будут порождаться правила, но и принципы установления синонимичных средств языка, что позволяет использовать новый подход на основе полей функционального переноса. В настоящий момент разрабатывается расширенная спецификация грамматики, в которой задаются категориально-функциональные признаки языковых объектов и структур и уточняются их вероятностные расширения [11].

Продолжается работа по формированию коллекции параллельных текстов научных и патентных документов на русском, английском, французском, немецком языках, на основе которых разрабатываются блоки семантико-синтаксических правил с учетом синтаксической неоднозначности. В настоящее время для существующей экспериментальной версии лингвистического процессора на базе английского и русского языков разрабатывается модуль французско-русского (и обратного) перевода [2].

Архитектурные решения и разрабатываемое функциональное наполнение многоязычной лингвистической базы знаний делают ее актуальным и востребованным ресурсом для широкого круга специалистов в области компьютерной лингвистики и интеллектуальных технологий, а также позволят использовать в учебно-методических целях, для сопоставительного анализа языковых структур и разработки сред, обучающих языку. Планируется также создание рабочего места переводчика с возможностью формирования и использования архива «переводческой памяти».

Немаловажным требованием к внешним интерфейсам является интуитивная понятность, эргономичность и простота освоения и использования.

Программный комплекс лингвистической базы знаний будет включать в себя компоненты, обеспечивающие:

- загрузку и отображение на экране исходного текста. Подсистема обработки входного текста в общем случае обеспечивает ввод исходного текста, при необходимости его последовательную фрагментацию на логико-структурные блоки, удобные для последующей обработки, ведение фрагментированного текста и подачу фрагмента в подсистему распознавания;
- распознавание и отображение функционально-семантических и синтаксических структур исходного текста. Подсистема обеспечивает семантико-синтаксический анализ заданного фрагмента на основе комплекса фразовых структур и словаря входного языка и построение его формальной семантической структуры. Предусматриваются возможность коррекции пользователем получаемых результатов и выбор наиболее релевантной структуры;
- построение и отображение функционально-семантических и синтаксических структур результирующего (их) текста (ов). Подсистема на основе формальной семантической структуры входного фрагмента текста, комплекса фразовых структур целевого языка (в общем случае – целевых языков) и словаря строит формальную семантическую структуру для выходного фрагмента текста. В связи с многовариантностью трансфера возможно построение набора выходных структур, характеризующихся как разной частотой использования в целевом языке, так и контекстом использования;
- создание, загрузку и отображение результирующего текста или результирующих текстов на нескольких языках в зависимости от режимов многоязычности и использования ранее созданных коллекций;
- поиск подходящих к исходному тексту фрагментов в архиве «переводческой памяти» и их повторное использование;
- сохранение готовых переводов в архиве «переводческой памяти» и в выходных файлах;
- отображение словарных статей, соответствующих лексике обрабатываемого текста;
- функционирование панелей настройки, управления и навигации;
- сохранение текущего состояния сеанса работы с системой, ведение истории сеанса и истории работы с текстом, чтобы при необходимости выполнить возврат к предыдущему варианту перевода;
- управление профилями и настройками пользователей;
- использование словарей – их подключение и удаление, поиск в них требуемой информации, извлечение выбранного варианта в рабочую область;
- управление коллекцией параллельных текстов, словарями, архивами «переводческой памяти», профилей и историй пользователей, служебными данными.

Принципиальная схема взаимодействия основных функциональных систем ядра многоязычной лингвистической базы знаний приведена на рисунке 1.

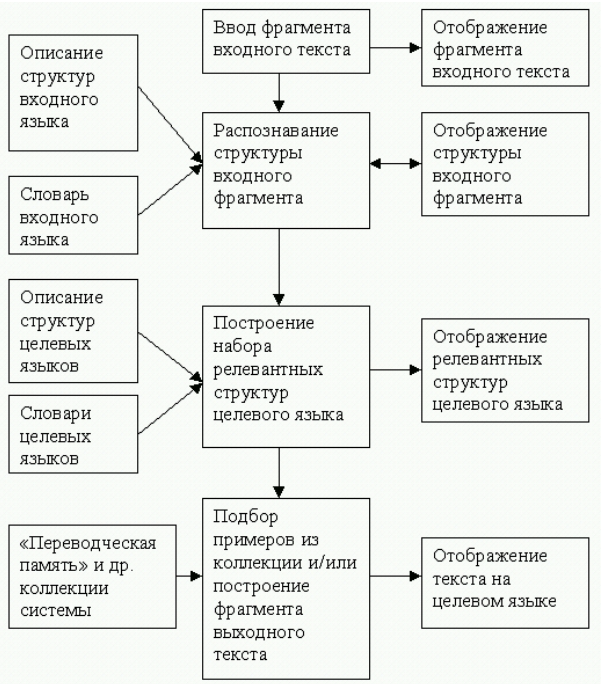

Рисунок 1. Функциональная схема ядра многоязычной лингвистической базы знаний

Описания структур входного и целевых языков представляют собой независимые комплексы иерархически выстроенных правил фразовых структур соответствующих языков, выстроенные в единой нотации. Внутренняя структура описания языка строится в виде сложно связанной таблицы (наподобие структуры данных в базе WordNet), каждая строка которой представляет собой правило в комплекте со ссылками-связями на родственные строки таблицы. Между собой комплексы фразовых структур разных языков связаны системой межязыковых отсылок, устанавливающих возможные направления трансфера от одного языка к другому (подобно межязыковому индексу в системе данных EuroWordNet). Подсистема фразовых структур языков обеспечивает ведение соответствующих архивов структур и вызов в подсистемы распознавания и трансфера необходимых пользователю комплексов структур языка. В основу иерархии фразовых структур положена когнитивная трансферная грамматика [5,9].

3. Данные и метаданные базы знаний

Данные, используемые программным комплексом лингвистической базы знаний, представляют собой набор разнородных массивов информации, хранящейся в различных форматах. Часть данных, например, коллекция параллельных текстов, допускает только пополнение коллекции, тогда как другая часть, например файлы конфигурации или истории работы пользователей, регулярно расширяются, обновляются и изменяются. Предполагается хранение части данных, требующих больших объемов памяти, в виде файлового архива сложной структуры с обеспечением его описания в базе данных в целях быстрого поиска.

Многоязычная лингвистическая база знаний использует следующие виды данных:

- коллекции параллельных текстов научных (статьи из разных областей науки) и патентных документов на русском, английском, французском языках в различных форматах хранения – doc, rtf, txt, pdf, html и др.;
- коллекция текстов «переводческой памяти»;
- комплексы фразовых структур, образующих иерархические системы правил для каждого обрабатываемого языка;
- словари;
- обрабатываемые системой тексты и их история;
- файлы конфигурации, профили и истории пользователей системы, описание файлового архива и прочие служебные файлы.

Метаданные описания файлового архива многоязычной лингвистической базы знаний должны полностью описывать структуру файлового архива – имя и тип элемента данных, тип коллекции, адрес размещения и объем, права доступа разных классов пользователей, время создания и обновления, язык, тематическую принадлежность и текущее состояние. Например, словари или архив «переводческой памяти» должны быть доступны для использования любому пользователю системы, тогда как загрузка параллельного текста в архив или модификация конфигурационных файлов доступны только специально уполномоченному пользователю лингвистической базы знаний или администратору.

Метаданные параллельных текстов включают также тип текста, формат и информацию о парном к данному тексте.

Элементы «переводческой памяти» кроме того характеризуются «автором» (пользователем, сохранившим текст) и целевым языком.

Словари, используемые в базе знаний, в общем случае внешние по отношению к системе. Метаданные словарей включают, кроме того, информацию о формате, авторе, языках исходном и целевом, тематике и лексическом объеме.

Метаданные комплексов фразовых структур содержат поля, однозначно характеризующие язык, иерархию его структур и порядок использования данных комплексов программными модулями лингвистической базы знаний.

Метаданные пользовательских текстов и история их обработки дополнительно включают информацию о пользователе, о структуре исходного текста и его сегментации, и последовательности произведенных над ним действий.

4. Заключение

Необходимость продвижения в решении проблемы создания систем машинного перевода и обработки знаний, содержащихся в текстах на различных естественных языках, настоятельно

диктуется интенсивностью и объемом межъязыковых и межкультурных коммуникаций в современный период, имеющих целью встраивание российского научного, политического и экономического сообществ в контекст общеевропейского и общемирового процесса развития.

В работе приводится описание основных архитектурных решений, данных и метаданных в многоязычной лингвистической базе знаний. Данная система базируется на сочетании методов структурно-семантического анализа фразовых структур и статистических характеристиках языковых объектов и создается на основе нового лингвистического ресурса «Полиглот» [4] для отладки семантико-синтаксических представлений в лингвистических процессорах систем машинного перевода и обработки текстовых знаний. Новая база знаний предназначена для широкого круга специалистов в области компьютерной лингвистики и интеллектуальных технологий в учебно-методических целях, для сопоставительного анализа языковых структур и разработки сред, обучающих языку. В рамках комплекса планируется создание рабочего места переводчика и формирование компоненты «переводческой памяти». Особое значение предлагаемая работа имеет для решения проблемы структурного анализа и компьютерного моделирования полнотекстовых научных и патентных документов.

Литература

[1] Азарова И.В., Митрофанова О.А., Синопальникова А.А. Компьютерный тезаурус русского языка типа WordNet // Труды международной конференции Диалог'2003 "Компьютерная лингвистика и интеллектуальные технологии", (Протвино, 11-16 июня 2003 г.) М., 2003, с. 43-50.

[2] Галина И.В. Вопросы функционально-семантической синонимии и трансформаций именных структур при автоматическом переводе (на материале французско-русских параллельных текстов научной тематики) // Труды международнойконференции Диалог'2006, Бекасово, 31 мая – 4 июня, 2006, М: Изд-во РГГУ, 2006, стр. 105 – 109.

[3] Добров Б.В., Лукашевич Н.В. Онтологии для автоматической обработки текстов: Описание понятий и лексических значений // Компьютерная лингвистика и интеллектуальные технологии: Тр. междунар. конференции Диалог'2006, Бекасово, 31 мая – 4 июня 2006 г., 2006, стр. 138-142.

[4] Козеренко Е.Б. Лингвистические аспекты информатики // Системы и средства информатики. Специальный выпуск Научно-методологические проблемы информатики. Москва, ИПИРАН, 2006, 88-111.

[5] Козеренко Е.Б. Моделирование переноса функциональных значений для англо-русского машинного перевода. // Труды Международной конференции Диалог'2004 "Компьютерная лингвистика и интеллектуальные технологии", М.: "Наука", 2004.

[6] Система визуальной локализации Alchemy Catalyst: http://www.alchemysoftware.ie/index.html

[7] EuroWordNet: Building a multilingual database with wordnets for several European languages. http://www.illc.uva.nl/EuroWordNet

[8] Ferrucci, D. and Lally, A. UIMA: an architectural approach to unstructured information processing in the corporate research environment // Natural Language Engineering 10 (3/4), 2004, 327–348.

[9] Kozerenko E.B. Cognitive Approach to Language Structure Segmentation for Machine Translation Algorithms // Proceedings of the International Conference on Machine Learning, Models, Technologies and Applications, June, 23-26, 2003, Las Vegas, USA.// CSREA Press, 2003. P. 49-55.

[10] Kozerenko E.B. INTERTEXT: A Multilingual Knowledge Base for Machine Translation // Proceedings of the International Conference on Machine Learning, Models, Technologies and Applications, June, 25-28, 2007, Las Vegas, USA.// CSREA Press, pp. 238 - 243, 2007.

[11] Kozerenko E.B. Semantic Approach to Language Structures Presentation for Machine Learning Algorithms Design. Proceedings of MLMTA'06. Las Vegas, June 26 – 29, 2006, - CRSEA Press, 2006, pp. 10-16.

[12] Pollard, C. and Sag, I.A. Head-Driven Phrase Structure Grammar. Chicago:University of Chicago Press, 1994.

[13] WordNet: a lexical database for the English language. http://wordnet.princeton.edu

MULTILINGUAL LINGUISTIC KNOWLEDGE BASE: ARCHITECTURE AND METADATA

N.V. Luneva

The paper describes some principal architectural decisions and types of metadata in the multilingual linguistic knowledge base founded on the new linguistic resource. The linguistic knowledge base is aimed at debugging semantic-syntactical representations in language processors of machine translation and text knowledge processing systems. The new knowledge base is being designed as a major test bed for the research community in the field of computational linguistics and intellectual technologies as well as for educational purposes, for comparative analysis of language structures and creating language training environments. The knowledge base features the component of the multilingual translation memory.