

Semp-L — программная среда для создания лингвистических проектов

Загорулько Ю.А., Кононенко И.С., Сидорова Е.А.

Российский НИИ Искусственного Интеллекта, Новосибирск
<http://www.riai.org.ru>

irina@mail.nsk.ru, {,zagor,sidorov}@iis.nsk.su

Аннотация.

В данной работе описывается программная среда *Semp-L*, предназначенная для поддержки работы лингвиста по созданию программ, решающих различные задачи, связанные с обработкой текстов. Кратко рассматриваются такие аспекты, как представление лингвистической информации в виде иерархии классов и системы продукционных правил, структура программной среды и предоставляемые ею возможности. Важной особенностью программной среды является наличие удобных средств для работы с банком текстов и словарем.

Введение

В Российском институте искусственного интеллекта активно ведутся исследования по созданию модели понимания текстов в ограниченной предметной области и построению на ее основе лингвистических процессоров и других прикладных систем [1,2,3]. В настоящее время возникла потребность в средстве, которое позволяло бы продвинутому пользователю (лингвисту-конструктору) создавать программные проекты (Л-проекты), обладающие следующими свойствами:

- ü Возможность свободного конструирования собственной иерархии понятий и/или использование уже существующей.
- ü Словарь, согласованный с существующей, динамически изменяемой иерархией понятий.
- ü Возможность модификации алгоритмов лингвистического анализа.

• Банк текстов.

• Современный, дружелюбный интерфейс.

По замыслу разработчиков программная среда *Semp-L* является *рабочим местом лингвиста*. Она должна обеспечивать поддержку распознавания текстов на естественном языке на основе семантически-ориентированного подхода [4], согласно которому понимание текста основано прежде всего на содержательной информации, характеризующей данную предметную область. Среда *Semp-L* создается на базе программного комплекса Semp-Taο [5], обеспечивающего пользователя гибкими средствами представления и обработки знаний.

1. Основные понятия и структура системы Semp-L

Программная обстановка *Semp-L* представляет собой многооконную пользовательскую оболочку, предназначенную для конструирования, настройки и отладки Л-проектов. Построение Л-проекта опирается на следующие основные понятия:

- 1) Иерархия понятий – иерархически упорядоченный набор классов, соответствующих выделенным конструктором лингвистическим понятиям и концептам предметной области с экспликацией их взаимосвязей. Иерархия отражает, с одной стороны, структуру конкретной предметной области (классы *объект*, *признак*, *значение*), а с другой стороны, - особенности структуры корпуса текстов и соответствующего подхода к анализу. Так, иерархия классов может содержать лексические, структурно-текстовые, тематические и грамматические понятия (*слово*, *тема*, *сегмент*, *существительное*). Классы определяются лингвистом непосредственно путем описания их свойств на некотором формальном языке и/или с помощью наследования свойств от ранее определенных более общих классов.
- 2) Словарь – список экземпляров классов, наследующих информацию об основе и флексии от базового класса *слово*, что обеспечивает словарный поиск. Словарь поддерживает возможности сортировки слов по классу, по значению какого-либо внутреннего элемента, обеспечивает создание подсловарей.
- 3) Лингвистический анализ – набор правил, осуществляющий необходимые процедуры обработки текстов: лексический, тематический, синтаксический, семантический анализ, генерацию выходного представления.
- 4) Банк текстов – набор анализируемых текстов, из которого выбирается тест для исполнения. Банк текстов поддерживает возможности поиска текста по признаку, сортировки текстов и создания подмножества текстов.

Создание Л-проекта включает в себя следующие этапы:

- описание понятий;
- создание и наполнение нового или подключение уже существующего словаря;

- .. создание или подключение уже существующего банка текстов;
- .. создание производционных правил;
- .. компиляцию всех созданных модулей;
- .. отладку Л-проекта.

Реализация перечисленных действий по построению лингвистического процессора осуществляется посредством пользовательских панелей. Главная панель возникает при вызове Semp-L и постоянно присутствует. Остальные вызываются с помощью главного меню или панели инструментов. Одна из них называется “Словарь” и предназначена для работы со словарем, другая — “Тексты” — предназначена для работы с текстами и выбора теста при отладке программы.

2. Описание оболочки Semp-L

Технологически система состоит из следующих частей:

- программная среда Semp-TAO, служащая для низкоуровневого программирования и отладки;
- визуализатор иерархии понятий, служащий для просмотра и редактирования иерархии понятий;
- словарь;
- наборы специализированных производций;
- банк текстов.

2.1. Иерархия понятий

Понятие может быть создано с помощью *визуализатора* иерархии понятий – специализированного средства визуального программирования – либо задано в текстовом виде, как описание класса среды Semp-TAO. Визуализатор отображает граф понятий, вводит структуризацию понятий по группам (слоям) и предоставляет средства для визуального редактирования иерархии понятий.

2.2. Словарь

Словарь является отдельным модулем системы Semp-L, который позволяет:

- сохранять и загружать словарь (согласованный с иерархией понятий);
- добавлять, удалять и редактировать слова;
- просматривать словарь по разным фильтрам;
- осуществлять поиск слова по основе;
- сортировать слова по значению конкретного внутреннего параметра;
- создавать подсловарь из списка специальным образом выбранных слов, что позволяет для каждой конкретной задачи использовать подсловарь с собственной тематикой.

На рисунке демонстрируется возможность автоматической выборки слов из словаря: список слов формируется для указанного класса и значения его слота.

2.3. Продукционные правила

Анализ текста осуществляется системой продукционных правил Semp-ТАО на основе иерархии понятий и словаря. Система позволяет использовать как стандартные наборы правил, так и создавать новые с помощью встроенного набора специализированных функций. Л-проект осуществляет следующие операции по обработке входного текста:

- лексический анализ - предварительный этап обработки, обеспечивающий преобразование входного текста в лексическую цепочку экземпляров классов,
- семантический или синтактико-семантический анализ - главный этап (по сложности и объему), на котором лексическая цепочка трансформируется в семантическую сеть, эксплицирующую содержание входного текста,
- генерацию выходного представления в требуемом формате (например, фрейм-структура или формат объектно-ориентированной базы данных).

2.4. Банк текстов

Банк текстов является отдельным модулем системы *Semp-L*, который обеспечивает работу лингвиста с репрезентативной выборкой из корпуса текстов. Тексты должны достаточно полно покрывать все возможные сочетания смыслов и их естественно-языковые способы выражения, а также проверять не только каждую группу правил, но и как можно больше их комбинаций. Банк текстов позволяет:

- сохранять и загружать тексты;

- просматривать, добавлять, удалять и редактировать тексты;
- классифицировать тексты с точки зрения различных релевантных для анализа признаков, таких как наличие однородных структур, отрицания и т.п.
- сортировать по признакам (группам признаков) и создавать подмножества текстов;
- выбирать тест для обработки.

Еще одним важным свойством банка текстов является взаимодействие со словарем с целью его пополнения при обновлении банка текстов.

2.5. Отладка

Используются стандартные методы отладки, подобные тем, что применяются в широко распространенных средах программирования. Отладочный интерфейс обеспечивает лингвисту возможность тестирования и контроля работы всех систем при запуске отдельных тестов. Так, имеется несколько вариантов экспликации текущего состояния рабочей сети или ее фрагмента, например, визуализация рабочей сети по мере исполнения того или иного содержательного этапа анализа (лексического, тематического).

3. Реализация

Среда *Semp-L* конструируется на базе программного комплекса Semp-TAO, который предназначен для создания широкого класса прикладных интеллектуальных систем, в том числе и таких, в которых требуется описание сложных по структуре и семантике предметных областей и сочетание логического вывода и вычислений над неточно заданными значениями.

Основной структурой данных в модели среды Semp-TAO является семантическая сеть, состоящая из объектов, связанных бинарными отношениями. Semp-TAO включает объектно-ориентированный язык представления и обработки знаний, который позволяет представлять понятия предметной области в виде классов объектов и отношений и специфицировать требуемые в приложениях процессы вывода и обработки данных в терминах продукционных правил.

Программная среда *Semp-L* реализуется в Windows 95/98/NT на языке C++ с использованием современных технологий создания пользовательских приложений (C++Builder).

Заключение

В данной работе описывается программная среда *Semp-L*, предназначенная для автоматизации работ лингвиста по созданию программ, решающих различные задачи, связанные с обработкой текстов. Кратко рассматриваются такие аспекты, как представление лингвистической информации в виде иерархии классов и системы продукционных правил, структура и предоставляемые ею возможности.

В настоящее время ведутся исследования по расширению возможностей программной среды. Предусматривается подключение грамматического словаря с соответствующей модификацией модуля дефиниций - интеграцией грамматической иерархии классов с определенными на ней морфологическими слотами. Планируется создание структур для сборки словокомплексов, развитие средств работы в режиме диалога.

Литература

Kononenko I., Sharoff S. Understanding Short Texts with Integration of Knowledge Representation Methods. In: Perspectives of System Informatics, Lecture Notes in Computer Science; Vol. 1181, Springer, 1996, pp. 111-121.

Нариньяни А.С. Автоматическое понимание текста - новая перспектива // Труды международного семинара Диалог'97 по компьютерной лингвистике и ее приложениям. - Москва, 1997, с.203-208.

Загорулько, И. Кононенко, И. Попов. Экспериментальная система понимания метеорологических телеграмм // Труды Международного семинара Диалог'99 по компьютерной лингвистике и ее приложениям. – Таруса, 1999.

Narin'yani A.S. Interaction with a limited object domain - ZAPSIB project. // Proc. of the int. conf. "Computational Linguistics-1980". -Tokyo, 1980.

Загорулько, И.Г. Попов. Описание сложных предметных областей на основе интеграции средств представления знаний. // Труды международного семинара Диалог'97 по компьютерной лингвистике и ее приложениям. -Москва, 1997. -С.110-115.