Инструментальная среда экспертной обработки японских текстов “NIHONGO”

Жалыбин П. П.
pjalybin@mail.ru

Мальковский М. Г.
malk@cs.msu.su

Факультет вычислительной математики и кибернетики МГУ им. М.В. Ломоносова

Инструментальная среда “NIHONGO” предназначена для автоматизации труда эксперта по работе с японоязычным лингвистическим материалом. Разрабатываемая система предоставляет пользователю средства для редактирования и лингвистического анализа японского текста, а также позволяет получать различную справочную информацию о строении японского языка. Система должна выполнять функции автоматизированного рабочего места лингвиста-япониста. Предполагается, что конечным пользователем системы будет исследователь японского языка (прежде всего), а также любой человек, сталкивающийся с японским текстом: переводчик, преподаватель, студент. Работа по созданию инструментальной среды “NIHONGO” [1] ведется на факультете ВМиК МГУ в сотрудничестве с Институтом востоковедения РАН.

В основе реализации лингвистического обеспечения разрабатываемой инструментальной среды лежит сущностный подход к языку, автором которого является З.М. Шаляпина [3, 4]. В рамках этого подхода единицы языка рассматриваются как сущности, являющиеся совокупностью присущих им свойств. Свойством сущности может быть любая значимая информация о ней, например, графическое написание, сведения о сочетаемости, семантическая классификация и т.д. Эти свойства (или в другой терминологии валентности) могут быть присущи непосредственно самой сущности, или могут наследоваться ею от более общих, базовых сущностей. Свойством сущности может быть любая значимая информация о ней, например, графическое написание, сведения о сочетаемости, семантическая классификация и т. д. Применение сущностного подхода к описанию естественного языка позволяет взглянуть на него как на иерархическую структуру сущностей: от отдельных морфем до частей речи и семантических классов. Другими словами, используя этот подход, можно описать естественный язык в стиле объектно-ориентированного программирования. Таким образом, авторами было решено для организации лингвистического обеспечения системы разработать специальный объектно-ориентированный язык LEELA[1], единый для описания всей информации о естественном языке, будь то грамматика, словарь лексем или семантическая классификация. Так как для приемлемой скорости работы информация хранится в бинарном виде, в состав системы включены средства для компиляции структур данных, которые закодированы на этом языке.

Инструменты работы с японским текстом.

Система поддерживает большинство из используемых кодировок японских символов (Unicode, JIS и др.) Пользователь системы может работать с японскими документами в обычном текстовом формате в любой поддерживаемой кодировке, а также с RTF файлами.

Разрабатываемая система предоставляет пользователю целый ряд удобных способов ввода японского иероглифического текста. Вводимая латинская или русская транскрипция японских слов автоматически преобразуется встроенными в систему средствами в японскую слоговую азбуку и далее в иероглифы. Система позволяет пользователю быстро найти нужный ему иероглифический знак. Кроме широко распространенных классических способов поиска

[1] List Entity Encoding LAnguage, Списковый язык кодирования сущностей

иероглифа по ключу, количеству черт, чтению, в системе “NIHONGO” также реализована возможность отыскания незнакомого иероглифа по составляющим его элементам. В статьи иероглифического словаря, который входит в лингвистическое обеспечение системы, включены сведения о простых элементах, составляющий каждый иероглиф. Пользователю, задавшему некое множество подобных элементов, система предложит ряд иероглифов, в состав которых входят указанные части. Таким образом, иероглифический ввод незнакомых иероглифов существенно упрощается.

Инструменты лингвистического анализа.

Японоязычный текст представляет собой непрерывную последовательность символов японской графики, разделяемых лишь знаками препинания. Поэтому важным инструментом лингвистического анализа японского текста, предоставляемым системой “NIHONGO”, является автоматическая сегментация, то есть выделение отдельных графико-морфологических составляющих. Процесс сегментации в системе совмещен с морфологическим анализом. В состав системы включен словообразовательный компонент, поэтому она приспособлена к работе с неопознанными словами, которые образованы путем сложения основ, конверсией частей речи, с помощью аффиксов и т.д. При наличии неоднозначности в процессе сегментации системой предлагается список вариантов разбиения сомнительного участка текста, отсортированный начиная с самого вероятного. Пользователь может выбрать правильный на его взгляд вариант сегментации или задать свой собственный вариант. При обнаружении неизвестных слов пользователю предлагается пополнить словарь, используя для этого интерфейс разработки сущностей.

Пользователю предоставлено два способа вносить изменения и дополнения в лингвистическое обеспечение системы. Можно непосредственно самому написать код на языке LEELA или использовать для этих целей удобный графический интерфейс разработки языковых сущностей. Используя этот интерфейс, пользователь может работать со свойствами сущности. При этом пользователь может узнать: является ли свойство наследуемым от более общей сущности или оно переопределено (или впервые указано) в описании данной сущности. Иерархическая структура сущностей отображается в виде графа, пользователь может выбрать нужную ему сущность, отследить ее наследников и общие сущности, наследницей свойств которых эта сущность является.

После сегментации текста на отдельные графико-морфологические элементы, система сопоставляет каждому элементу определенную сущность. Этой сущностью может быть отдельная морфема при однозначном определении, набор морфем-омографов, или же, например, если система встретила в тексте незнакомое слово с морфологическими свойствами глагола, то этой сущностью в этом случае может быть глагол как часть речи. Пользователь может скорректировать предложенный системой вариант сопоставления отрезков текста сущностям. При этом системой ведется статистика таких сопоставлений для обеспечения более корректной работы в будущем.

Далее текст рассматривается системой уже как последовательность сущностей, всей совокупности их свойств. Синтаксический анализ рассматривается как процесс заполнения актантных валентностей [5], которые присущи каждой сущности в этой последовательности. Причем сущность может реализовывать свою собственную валентность, а может наследовать ее от другой сущности, синтаксически связанной с ней (как, например, предлог наследует валентность своего хозяина [6]). При описании валентности, характерной для некоторой сущности, указывается список сущностей, которыми эта валентность может заполняться. На заполнение валентности могут накладываться различные ограничения: семантические, синтаксические, лексические и другие.

Синтаксическая структура предложения отображается системой в виде графа. Вершинами этого графа являются сущности, а дугами – заполняемые синтаксические валентности или отношения кореферентности. Система в состоянии автоматически заполнить некоторые валентности, однако пользователь всегда может исправить предложенный системой вариант. При появлении противоречий между предложенным пользователем заполнением

валентности и описанными ограничениями на ее заполнение система может предложить пользователю снять это противоречие, исправив свою ошибку, или изменить множество сущностей – допустимых заполнителей валентности. Системой ведется статистика использования той или иной сущности в качестве заполнителя валентности сущностей. Это позволяет выявить часто встречающиеся синтаксические конструкции и, тем самым, повысить эффективность синтаксического анализа.

Организация лингвистического обеспечения системы и язык его описания.

Хотя лингвистическое обеспечение системы представляет собой связную иерархическую структуру сущностей, все же можно разделить его на следующие составные части. Это иероглифический словарь, содержащий информацию о сущностях – символах японской графики. Грамматический и семантический словари, которые хранят данные о неких обобщенных сущностях языка: части речи, семантические классы. Словарь японских элементарных языковых единиц: лексем, корней, аффиксов, служебных слов, который отражает минимальные сущности в структуре языка.

Для кодирования всех перечисленных словарей используется специальный объектно-ориентированный язык представления данных – LEELA. Кратко отметим его основные особенности. Этот язык имеет списковую структуру, подобно языку LISP. Код на языке LEELA представляет собой список сущностей. Для каждой сущности обязательно указывается ее уникальное имя; список сущностей-родителей, свойства которых наследуются; список свойств сущности. Свойство сущности состоит из уникального имени и произвольной многоуровневой списковой структуры. При построении полного набора свойств сущности, в него добавляются все свойства сущностей-предков, причем свойства с одинаковыми уникальными именами переопределяют друг друга. Однако, используя специальные ключевые слова, можно дополнить или удалить наследуемые свойства. Также для описания списка сущностей в языке предусмотрена возможность использования логических операторов пересечения и отрицания сущностей как множества свойств (операторы & и ~). Наследование сущностями-потомками свойств базовых сущностей-родителей позволяет подразумевать под более общей сущностью любую сущность из всей совокупности ее потомков. Таким образом, язык LEELA оказывается удобным для представления лингвистической структуры естественного языка в рамках сущностного подхода.

Словарно-справочные инструменты системы.

Многообразие информации, составляющей лингвистическое обеспечение системы, делает возможным использовать систему в качестве эффективного инструмента по получению справочной информации о различных сторонах японского языка. Система предоставляет пользователю информацию о лексике японского языка, с ней возможно работать как с удобным электронным японско-русско-английским словарем. Можно получить грамматические характеристики лексемы, для спрягаемых частей речи можно узнать парадигму спряжения. Используя словообразовательный компонент, пользователь может получить информацию об основах и аффиксах, составляющих сложное слово. Для проведения лингвистических экспериментов можно получить статистику использования слов в различных контекстах, а также частоту употребления. Представляется возможным посмотреть список синтаксических валентностей, присущих определенной единице японского языка. Система обладает функциями словаря синонимов и лексико-семантического словаря. Системой предоставляются обширные сведения о японской графике: чтение иероглифов, их написание, смысл, частота, с которой они встречаются, количество черт, ключ, а также список элементов, составляющих иероглиф. Объем и многосторонность предоставляемой справочной лингвистической информации открывают перспективу широкого практического применения системы “NIHONGO”.

Заключение.

Инструментальная среда “NIHONGO” представляет собой действующую компьютерную систему для работы с японским лингвистическим материалом. Теоретической основой

реализации системы “NIHONGO” (как и комплекса ЯРАП/1 [2]) является сущностный подход к языку. Для описания лингвистической информации авторами доклада разработан объектно-ориентированный язык LEELA. Система предоставляет эксперту возможность работы с японской графикой, морфологией, синтаксисом и другими аспектами японского языка. Так, система обладает способностью предоставлять практически необходимую при работе с японским текстом справочную информацию. Функциональные особенности среды “NIHONGO” позволяют использовать ее как для экспериментальных лингвистических исследований, так и для чисто прикладных целей.

Литература.

1. П.П. Жалыбин. Японоязычный лингвистический процессор “NIHONGO”. – В кн.: Сборник тезисов лучших дипломных работ 2004г. – М.,МГУ, 2004, с. 87-88.
2. З.М.Шаляпина, Л.С.Модина, М.И.Канович, В.И.Любченко, А.С.Панина, Н.И.Сенина, В.И.Сивцева, Е.С.Тарасова, И.М.Хайлова, О.А.Штернова. Экспериментальный комплекс ЯРАП для лингвистических исследований в области японско-русского автоматического перевода: первая очередь. Москва, ИНИОН РАН, 2001.
3. З.М.Шаляпина. Оппозиция "часть-целое" и сущностный подход к моделированию языковой компетенции. – В кн.: Роман Якобсон: тексты, документы, исследования. – М.: РГГУ, 1999, с. 541-551.
4. Л.С.Модина, З.М.Шаляпина. Принципы организации лингвистических знаний в объектно-ориентированной модели лексико-морфологической системы японского языка. – В кн.: DIALOG '95. Труды Международного семинара по компьютерной лингвистике и ее приложениям. – Казань, 1995, с. 198-205.
5. Мальковский М.Г. Диалог с системой искусственного интеллекта. М.: МГУ, 1985.
6. А.С.Панина. К проблеме описания служебных единиц (на японском материале). – В кн.: DIALOG '2004. Труды Международного семинара по компьютерной лингвистике и ее приложениям. с. 487-492 .