База знаний интеллектуальной справочной системы по русскому языку

Русецкий К.В.
Кафедра интеллектуальных информационных технологий
Белорусский государственный университет информатики и радиоэлектроники
Минск, Республика Беларусь
e-mail: rusetski.k@gmail.com

***Аннотация*—Описан подход к построению лингвистической базы знаний по русскому языку на базе открытой семантической технологии, развиваемой в рамках международного проекта OSTIS**

Ключевые слова: интеллектуальные системы; русский язык; лингвистика; базы знаний

I. Введение

Стремительное развитие и все более глубокое проникновение компьютерных систем в самые разнообразные сферы деятельности человека вызывает необходимость в снижении входного порога для новых пользователей и, как следствие, уменьшении расходов на их подготовку. В этом смысле наиболее привлекательным видится использование привычного для пользователя языка для организации его диалога с компьютерной системой. Для этого необходима реализация естественно-языкового пользовательского интерфейса (ЕЯ-интерфейса). ЕЯ-интерфейс обладает рядом преимуществ:

- Для работы с системой необходима лишь минимальная подготовка, выражающаяся в уверенном владении языком
- Естественный язык позволяет пользователю просто и быстро задавать произвольные запросы к системе.

ЕЯ-интерфейс наиболее хорошо подходит для интеллектуальных систем в силу того, что взаимодействие на естественном языке само по себе является интеллектуальной задачей [7]. Чтобы интеллектуальная система была способна общаться с пользователем на его языке, она должна обладать знаниями об этом языке. Такой информацией ее обеспечит лингвистическая база знаний.

II. Общая структура интеллектуальной справочной системы

Интеллектуальная справочная система состоит из трех основных компонентов:

1. База знаний [2], представленная в виде семантической сети [6].
2. Машина обработки знаний, представленная агентами обработки семантической сети.
3. Пользовательский интерфейс использует для общения с пользователем подмножество SCg языков [5].

В интеллектуальных системах [2] информация представляется в виде семантической сети, что позволяет оперировать не только фактографической информацией, но и осуществлять навигацию по установленным отношениям [3] в рамках предметной области прикладной вопросно-ответной системы.

Исходной формой представления знаний является текст на линейном графовом языке SCs [5], который для удобства восприятия может быть представлен в виде текста на языке SCn [5]. Пример подобного текста, описывающего понятие, входящее в лингвистическую базу знаний, приведен на рисунке 1. Графическое изображение фрагмента семантической сети, соответствующего описанию грамматической информации [4] о некотором слове, приведен на рисунке 2.

Знания о естественном языке, записанные в лингвистической базе знаний, используются для анализа естественно-языкового текста, а также для его синтеза. Лингвистическая база знаний может также использоваться в интеллектуальных обучающих системах для генерации заданий и их решения, для самоконтроля пользователя-обучающегося. Эта часть базы знаний может быть выделена в отдельный IP-компонент и может использоваться в качестве предметной базы знаний по русскому языку. Отметим также важность интеллектуальных вопросно-ответных систем, т.к. они составляют основу интеллектуальных систем.

имя существительное
= *существительное*
= *множество всех существительных*
= *множество всех имен существительных*
∈ *Лингвистика русского языка*
∈ *множество*
⊂ *часть речи*
⊂ *самостоятельная часть речи*
⊃ собирательное существительное
– Разбиение (по собственности-нарицательности):
  • собственное существительное
  • нарицательное существительное
– Разбиение (по одушевленности):
  • одушевленное существительное
  • неодушевленное существительное
– Разбиение (по образованию):
  • отглагольное существительное
  • отпричастное существительное
– Примеры:
  • *солнце*
  • *подснежник*

Рис. 1. Статья на псевдоестественном языке

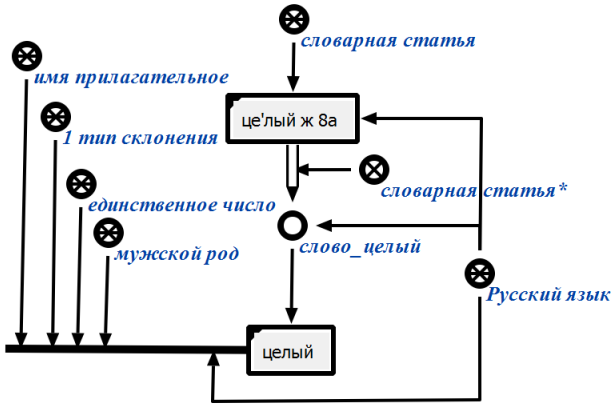

Рис. 2. Графическое представление фрагмента семантической сети.

III. Построение лингвистической базы знаний

Построение лингвистической базы знаний на основе семантической технологии OSTIS происходит в несколько этапов:

1. Составление тестового сборника вопросов.
2. Запись ответов на вопросы сборника
3. Выделение на основании полученных ответов основных понятий и отношений между ними.
4. Составление и запись статей на формальном языке для выделенных понятий, отношений и утверждений о них
5. Сборка базы знаний
6. Тестирование и верификация базы знаний на предмет ошибок и противоречий.

Одних лишь понятий и утверждений в лингвистической базе знаний, однако, недостаточно. Описание естественного языка будет недостаточно полным без описания его лексической и грамматической составляющих. Для этого на четвертом этапе на основании электронного варианта грамматического словаря А.А. Зализняка в автоматическом режиме по определенным правилам дополнительно генерируются тексты на формальном языке, соответствующие фрагментам семантической сети, подобным приведенному на рисунке 2.

Рассмотрим этапы подробнее на примере раздела «Части речи» лингвистической базы знаний.

A. Сборник вопросов

На данном этапе определяется перечень вопросов, ответы на которые должна обеспечивать база знаний. Примеры таких вопросов:

1. Что такое имя существительное?
2. Какие падежи в русском языке являются косвенными?
3. По каким признакам классифицируются существительные?

B. Ответы на вопросы

На данном этапе ответы на вопросы сборника записываются на естественном языке. Например, на вопросы предыдущего подраздела можно дать такие ответы:

1. Самостоятельная часть речи, объединяющая слова с грамматическим значением предметности.[4]
2. Родительный, дательный, винительный, творительный и предложный.
3. По одушевленности, собственности-нарицательности и по образованию.

C. Выделение понятий и отношений

На данном этапе из текстов ответов выделяются ключевые понятия и отношения. Из текстов ответов предыдущего подраздела можно выделить следующие понятия: (самостоятельная) часть речи, слово, грамматическое значение; родительный, дательный, винительный, творительный и предложный падежи; и др.

D. Запись на формальном языке

Выделенные на предыдущем этапе понятия записываются на формальном языке. Пример описания понятия на языке SCn приведен на рисунке 1. На этом этапе также описываются логические утверждения [1] о понятиях, например: «Для любого существительного pluralia tantum верно, что его полная парадигма состоит из шести падежных форм»

E. Сборка, тестирование, верификация

Полученные SCs-тексты передаются сборщику, который производит первоначальную проверку текстов и преобразовывает их в формат SC-хранилища, после чего осуществляется верификация на предмет логических противоречий, омонимичных идентификаторов и пр. Тестирование состоит в задании системе вопросов из тестового сборника и проверке правильности ответов на них.

IV. Заключение

Была рассмотрена структура лингвистической базы знаний, предложен подход к ее построению, описаны этапы его реализации на примере раздела лингвистики, описывающего части речи и их грамматические характеристики. Следует отметить, что лингвистическая база знаний может быть использована самостоятельно для получения информации об устройстве русского языка, или в составе обучающих, информационно-справочных систем, или в качестве основы для построения естественно-языковых человеко-машинных интерфейсов.

[1] Математическая логика: Учеб. пособие / Л.А.Латонин, Ю.А.Макаренков, В.В.Николаева, А.А.Столяр. Под общ.ред. А.А.Столяра. - Мн.: Выш. школа, 1991. - 269с.

[2] Базы знаний интеллектуальных систем / Т.А. Гаврилова, В.Ф. Хорошевский – СПб: Питер, 2000. – 384с.

[3] Новиков Ф.А. Дискретная математика для программистов / Учебное пособие 2-е изд., перераб. и доп. - СПб.: Питер, 2003 г. - 364 с.

[4] Д.Э. Розенталь, И.Б. Голуб Русский язык / Учебное пособие 2-е изд., перераб. и доп. - М.: Рольф, 2001. - 382 с.

[5] Open Semantic Technology for Intelligent Systems[Электронный ресурс] / Ostis Минск, 2010 http://ostis.net

[6] Харари Ф. Теория графов. Пер. с англ. 3-е изд. - М.: КомКнига, 2006. - 296с.

[7] Общение с ЭВМ на естественном языке. Попов Э. В.— М.: Наука. Главная редакция физико-математической литературы, 1982.— 360 с.