Компьютерная система семантической классификации лексики

А.В. Рафаева e-mail: anna_raf@rambler.ru
Московский государственный университет

Цель описываемой компьютерной системы – предоставить пользователю возможность задавать, редактировать и анализировать семантические отношений между лексическими единицами в форме, ориентированной на компьютер и программную обработку, а не на человека. Лингвистическая постановка задачи принадлежит А.А. Кретову и описывается в [1].

В данной системе под лексической единицей (леммой) понимается произвольная последовательность символов, не содержащая цифр, зарезервированных служебных символов, а также начальных и концевых пробельных символов. С точки зрения пользователя, леммой является слово (реже словосочетание) естественного языка в словарной форме. Единицей словника является лемма со снятой полисемией и ононимией (лексико-семантический вариант, ЛСВ), представленная в виде тройки *<l, n, m>*, где *l* – лемма, *n* – номер омонима (натуральное число), *m* – номер значения (последовательность символов, начинающаяся с цифры).

На множестве таких троек определены отношения *тождества*, *предшествования* (при этом множество является строго упорядоченным) и *лексикографического совпадения* (в последнем случае учитываются только значения лемм без учета номера омонима и номера значения). Единицами словаря являются пары, в которых каждому ЛСВ соответствует словарная дефиниция, возможно, пустая.

Семантические отношения между элементами словаря описываются в виде последовательностей (цепочек) ЛСВ, каждая из которых представляет собой путь в ориентированном графе и строится вручную следующим образом:

1. Входная единица представлена в текстовом примере в виде одной из возможных словоформ. Пользователь отождествляет словоформу со словарной формой (леммой);
2. Лемме, служащей входом (входной лемме) ставится в соответствие одно из толкований, представленных в словаре дефинициями. Если нужное значение присутствует в базе данных системы, толкование и соответствующий ЛСВ может быть выбран из множества значений. В противном случае соответствующие данные вводятся вручную с клавиатуры. Таким образом выделяется ЛСВ, который и служит входным узлом цепочки;

3. В дефиниции выделяется метаслово, которое будет служить следующим узлом цепочки (входом следующего уровня);
4. Метаслово приводится к словарной форме, после чего процедура повторяется.
5. Сигналом достижения конца цепочки служит появление одного из ЛСВ, уже присутствующих в данной цепочке (т.е. данный ЛСВ в конечном итоге толкуется сам через себя, что приводит к появлению цикла), или достижение одного из ЛСВ, объявленных корневыми .
Например (в приведенном примере цифра перед словом обозначает номер омонима, цифра после слова – номер значения):
1стоять1 – 1нога1 – 1человек1 – 1существо2 – 1животное1 – 1существо1.

На множество цепочек и входящие в них ЛСВ накладываются следующие ограничения:

- Ограничение на толкование. Каждый ЛСВ, кроме концевого, должен иметь непустое толкование;
- Ограничение на длину цепочки. Цепочка должна содержать хотя бы два узла (т.е. появление изолированных узлов недопустимо), максимальная длина цепочки может быть ограничена пользователем. По умолчанию ограничение на максимальную длину цепочки отсутствует;
- Ограничение на согласованность. Множество цепочек должно быть согласованным, т.е. от каждого ЛСВ существует один и только один путь к конечному узлу, что соответствует единственности толкования ЛСВ, за исключением случаев толкования ЛСВ через себя (циклов);
- Ограничение на количество циклов. В цепочке не может присутствовать более одного цикла, т.е. допустимо появление не более одного ЛСВ, толкующегося через себя. Последнее требование по желанию пользователя может быть ужесточено с использованием отношения лексикографического совпадения вместо отношения тождества. В таком случае в цепочке не может появиться более двух одинаковых лемм, как в приведенном выше примере.

Указанные требования реализуются с помощью ряда фильтров, накладывающих ограничения на исходные данные, введенные пользователем.

Компьютерная система реализована в среде программирования C++ Builder с использованием библиотеки шаблонов STL, входящей в

комплект поставки. Система работает в среде Windows'98 и выше. Система разрабатывалась в несколько этапов. Задачей первого этапа была реализация средства для создания, редактирования и сохранения исходных файлов и построения на их основе словаря и цепочек толкований. На этом же этапе был разработан и модуль, обеспечивающий применение фильтров к исходным данным, а также ряд вспомогательных модулей. Таким образом, на первом этапе проектирования было разработано средство, позволяющее производить первичное накопление данных и проверять некоторые исследовательские предположения о виде семантических отношений между ЛСВ.

Однако опыт работы с системой, созданной на первом этапе проектирования, показал, что для пользователя составление цепочек толкований, ориентированных на компьютерную обработку, представляет собой в достаточной степени трудоемкую задачу. Кроме того, построение цепочек без вспомогательного компьютерного средства привело к появлению большого количества трудноуловимых ошибок ввода. Поэтому следующей задачей явилось создание модуля, позволяющего вводить данные в удобном для человека виде. В этом же модуле должна быть реализована система построения автоматических подсказок (автодополнения) в том случае, если вновь вводимая лемма уже присутствует в базе данных. Наконец, последняя задача второго этапа – реализовать возможность задания различных режимов анализа цепочек, к которым относятся:

- задание ограничений на длину цепочки;
- разрешение / запрет на появление циклов;
- функция понижения регистра текста в исходном файле,

а также некоторые другие.

В настоящее время этот модуль частично реализован и проходит тестирование.

Наконец, следующим этапом работы над системой должна явиться разработка средств анализа полученных данных, включающих, в частности, возможности сортировки, индексации и фильтрации цепочек, хранящихся в базе данных. Этот модуль пока не реализован, однако данные, полученные в результате работы системы, могут быть импортированы в программу MS Excel, что позволяет воспользоваться возможностями электронных таблиц для получения статистических сведений о полученных результатах и представления их в наглядном для пользователя виде.

В настоящее время система содержит следующие модули:

- Модуль, обеспечивающий пользовательский интерфейс и задание режимов работы системы;
- Модуль ввода, редактирования и сохранения исходных данных;
- Модуль поиска фрагмента текста (последовательности символов) как на множестве исходных, так и на множестве результирующих данных;
- Модуль, содержащий служебные классы и функции, описывающий отношения на множестве ЛСВ, ограничения на вид цепочек и т.п.;
- Модуль анализа построенных пользователем последовательностей толкований, включающий фильтры и обработку ошибок ввода;
- Модуль, обеспечивающий удобный для пользователя ввод данных и функцию автодополнения исходных данных.

Предполагается разработка дополнительных модулей, в задачи которых должны входить:

- Функции обработки, анализа и, возможно, визуализации базы данных системы;
- Функции работы с несколькими наборами данных, включая установки по умолчанию для каждого из наборов.

Последняя возможность позволит использовать систему не только для задания семантической классификации лексики, но, в частности, и для описания системы семантических оппозиций в фольклорном тексте, которые будут представлены уже не множеством лексических единиц, но множеством наборов таковых.

Литература

1. Кретов А.А., Рафаева А.В. К созданию компьютерной системы семантической классификации лексики (в печати).