

Разработка формальных спецификаций лексики научно-технических текстов для их анализа

и перевода

Федюрко В.В., Щоголева Н.Н.

Институт кибернетики имени В.М. Глушкова НАН Украины

В сообщении обсуждаются проблемы, связанные с составлением и верификацией формальных спецификаций лексики для двух систем DUET и FEST обработки научно-технических текстов на флективных языках - русском и украинском. Обе системы получены с помощью генератора языковых процессоров DUAL [1], что обеспечило единство словарных структур и независимость программных компонентов систем от конкретного наполнения этих структур (автономность словарей).

Первая из систем DUET [2] предназначена для перевода с русского языка на украинский с последующим постредактированием результата перевода. Благодаря автономности словарей система DUET оказалась пригодной для обратного перевода с украинского языка на русский. Другая система FEST [3] создана с целью определения тематики текстов, уровня их специализации и других свойств на основе частотных списков известных (найденных в словаре) и неизвестных (отсутствующих в словаре) слов и словосочетаний.

Накопление слов в словарях осуществляется с использованием текстовой идеологии. В процессе обработки текста к каждой словоформе применяется морфологический анализ, в результате которого окончание может быть найдено, а основа - нет, если она отсутствует в словаре. Такие словоформы заносятся в список ненайденных в виде двух частей - основы и окончания. После завершения обработки текста составляется частотный список словоформ, где все словоформы с общей основой представлены основой и списком различных окончаний словоформ. Этот список не всегда достоверный, но для отбора лексики в словарь вполне пригодный.

Пополнение словарей на основе частотного списка дает хорошие результаты. Добавление только наиболее употребительных слов при переводе значительно увеличивает объём текста, не требующего постредактирования. Введение же в словарь слов, употребляющихся не более 1-2 раз, операция более трудоемкая, чем их непосредственный перевод при постредактировании.

Лексика научно-технических текстов условно разделена на три части (типа): общеупотребительные слова, лексика некоторой области знаний и терминология узкой специальности данного автора или группы авторов. В зависимости от типа описание лексики делится на так называемые словарные модули. Модуль общеупотребительной лексики может служить основой для всех словарей и быть подготовлен заранее, а потом пополняться.

Спецификация лексики, содержащая информацию, необходимую для ее анализа и перевода, составляется на языке Dual. Язык рассчитан на знание русского и украинского языков в пределах средней школы. Спецификация - это последовательность правил, разделенных точкой с запятой. Приводим синтаксис часто употребляемого правила перевода:

<правило> ::= <вх. цепочка> => <вых. цепочка>:<схема перевода> ;

<схема перевода> ::= <левая часть> = <правая часть> ;

Входная цепочка - это либо целое слово, либо основа нескольких словоформ, либо словосочетание русского языка. Выходная цепочка означает то же самое на украинском языке. Левая часть схемы содержит информацию, необходимую для анализа входной цепочки, а правая - для синтеза выходной цепочки. Примеры правил можно найти в [1].

Составление спецификации - процесс творческий и вначале выполнялся вручную. Но при большом количестве неизвестных слов составление спецификаций становится трудоемкой работой. Рассмотрим некоторые способы автоматизированного построения спецификаций для распознавания словоформ.

1. Автоматическое построение правил. Частотный список дает не только основу словоформы, но и весь спектр окончаний словоформ, встретившихся в текстах. Для наиболее употребляемых словоформ набор окончаний может быть достаточно богатым, чтобы однозначно определить класс словоформ, а значит, и основную форму, что позволяет автоматически сформировать правило для анализа. Конечно, такой способ требует большой выборки текстов для составления частотного списка, проверки правильности составленного правила и дает лишь частичный результат.
2. Использование опыта. Накопленный в результате построения спецификаций опыт можно использовать при диалоговой обработке списка неизвестных слов. Около 80% работы по составлению спецификаций можно отнести к рутинной. Например, в украинском языке большинство имён прилагательных относится к твёрдой группе и склоняется по одному образцу. Диалоговыми средствами можно выделить группу слов, склоняющихся или спрягающихся по одному образцу и указать шаблон для правила. Тем самым одновременно можно построить правила для десятков лексем. Даже при ручной работе такой способ дает значительное ускорение.
3. Получение спецификации лексем для анализа из спецификации перевода. Большинство правил спецификации лексем для системы FEST (для анализа текстов) можно получить из правил перевода, отбросив правую часть схемы вместе со знаком равенства. Однако, не все правила можно получить таким способом.
4. Разработку спецификации перевода лучше начинать с разработки и тестирования спецификации для анализа. Единственный пока способ облегчения труда состоит в автоматическом получении спецификации обратного перевода. Но не каждое правило поддается конверсии, только простые, имеющие приведенный выше синтаксис. Но зато соблюдается принцип: человек избавляется от рутинной работы и занимается творческой.

Построение машинного словаря по спецификации выполняет общий для обеих систем генератор словарей Gen.

Спецификация - это текст на формальном языке и как таковой может содержать синтаксические и семантические ошибки. Поэтому при всех способах формирования спецификаций требуется проверка их правильности. Проверка синтаксической правильности осуществляется генератором Gen в режиме отключения действий по генерации словарей.

Затем происходит генерация машинного словаря. При пополнении уже существующего словаря могут появиться правила с одинаковыми основами, что запрещено. При этом

возможны три варианта: правила идентичные (одно из них исключается); одно из правил дает более богатый набор словоформ, тогда правило с меньшим набором исключается; правила дополняют друг друга, тогда выполняется их слияние. При кажущейся простоте эта работа тоже трудоёмка для больших спецификаций. Она облегчается упорядочиванием правил по алфавиту. Однако удобнее иметь возможность параллельного просмотра обоих правил, чтобы в режиме диалога выбрать нужный вариант или осуществить их слияние.

Построенный после этого словарь может содержать семантические ошибки, то-есть выполнять ошибочный анализ или синтез. С целью поиска таких ошибок предусмотрена генерация всех словоформ по построенному словарю и их проверка на принадлежность к соответствующему языку.

Словарные модули для различных специальностей одного и того же языка или нескольких языков составляют текстовую библиотеку генератора DUAL и могут быть

использованы в любой конфигурации, начинающейся модулем с общеупотребительной лексикой, для построения машинного словаря для отдельной специальности. В настоящее время составлены спецификации по отдельным разделам кибернетики, математики и законодательства.

Литература

Мищенко Н.М. О генерации языковых процессоров на основе формальной спецификации лексики обрабатываемых текстов. (Труды настоящего Семинара).

Мищенко Н.М. Мобильная система пословного перевода научных текстов на флективных с использованием формальных спецификаций //Кибернетика и системный анализ. - 1999. - N1. - С. 33-42.

Мищенко Н.М., Федюрко В.В., Щоголева Н.М. Мобильная система статистической обработки профессиональных текстов в целях определения их тематики // Проблемы программирования. - 1999. - N2. С. 11-18.