Структура и функции автоматического многоязычного переводного словаря

The Structure and Functions of the Automation Multilingual Translation Dictionary

Грязнухина Т.А.

Украинский языково-информационный фонд НАНУ

Рассматриваются вопросы разработки структуры многоязычного переводного словаря и его функциональные возможности, а также некоторые аспекты построения пользовательского интерфейса. Автоматический многоязычный переводной словарь предназначается для использования в автономном режиме (как информационно-справочная система для переводчика), а также для использования в контурах систем машинного перевода. При разработке структуры словаря учитывались требования, предъявляемые к программным продуктам подобного типа. Слова в словаре сопровождаются свойственными им морфологическими, синтаксическими и семантическими признаками, необходимыми для работы алгоритмов МП.

Автоматический многоязычный переводной словарь (АМПС) разрабатывается как часть интегрированной лексикографической базы Украинского языково-информационного фонда Национальной академии наук Украины (УЯИФ НАНУ) .

При разработке проекта Словаря определяющую роль играли выдвинутые перед ним задачи:

1) служить основным инструментом поиска (установления) лексических переводных эквивалентов (ЛПЭ) в системе Машинного перевода, создаваемой в УЯИФ НАНУ. Первая версия Системы ориентирована на машинный перевод научно-технических, социально-политических текстов с украинского языка на языки русский, английский, испанский, немецкий, французский и на перевод с указанных языков на украинский;

2) для работы в автономном режиме словарь должен быть интегрирован в общую лексикографическую информационную базу Фонда в качестве информационно-справочной ее компоненты, выполняющей функцию своеобразного АРМа человека-переводчика или пользователя, пишущего текст на одном из языков Словаря;

3) в системах автоматической обработки текстов (АОТ), написанных на любом из перечисленных языков, Словарь должен служить источником грамматической информации, необходимой для работы алгоритмов автоматического морфологического, контекстуального и синтаксического анализов (если речь идет об обработке русских или украинских текстов) или морфолого-синтаксического (при АОТ английских, испанских, немецких, французских), а также для алгоритмов лемматизации и синтеза. Последние обеспечивают возможность вхождения в словарь при выполнении любой из названных функций не только со словом в его исходной форме, но и непосредственно из текста (т.е. с любой парадигматической формой слова).

Выполнение названных задач обеспечивается специальной структурой АМПС. В словаре каждому из языков отводится отдельная зона, состоящая из унифицированных полей лексичекой, грамматической,

семантической и прагматической информации. Основными элементарными единицами лексических полей (ЛексП) являются лексы - слова в их конкретных значениях. Эксплицитным выражением значений лексов выступают номера лексов в ЛексП, совпадащие с номерами толкований соответствующих значений, фиксируемых в семантическом поле (СемП). Переводные эквиваленты во всех языковых зонах представлены семантическими группами (СГ), объединяющими лексы с одинаковым значением. СГ может состоять из одного элемента, если в языке для передачи данного значения многозначного слова отсутствуют синонимы. Всем элементам внутри одной СГ присваиваивается одинаковый номер толкования. Переводные эквиваленты конкретной СГ во всех языковых зонах словаря записаны под одним номером. Именно эта связь переводных эквивалентов в разных языках по признаку общности выражаемого ими значения обеспечивает возможность перевода слов с любого языка на любой из заданных в словаре.

Внутри семантических групп лексы проранжированы по степени удаленности их значений от толкования, присвоенного соответствующей СГ. При использовании Словаря в системе МП установленные ранги служат формальным средством для алгоритмического разграничения полисемантических и синонимических переводных эквивалентов: слово с рангом 1 получает статус лексического варианта перевода (ЛВП), остальные слова СГ образуют множество возможных синонимических замен этого варианта.

В случае выбора пользователем режима перевода текста без постредактирования результатов перевода с экрана компьютера средствами, предоставляемыми ему системой МП, информация о синонимах ПЭ будет потеряна. Если этап постредактирования результатов МП в диалоговом режиме не исключается из процесса перевода, пользователь помимо внесения грамматических правок в перевод может, открыв синонимы определенного ЛВП, заменить его (на свое усмотрение) любым другим словом из заданных синонимов в СГ. Замена происходит автоматически, поскольку программой перевода предусмотрено согласование форм синонимов с формой ЛВП.

Каждая словарная единица языковой зоны в АМПС сопровождается определенными морфологическими, синтаксическими, семантическими и прагматическими характеристиками, разнесенными по соответствующим полям под номером своего лекса. Для всех языков релевантными являются признаки подъязыка (или подъязыков), код грамматического класса и номер парадигматического класса, репрезентирующий словоизменительный тип слова.

Признак подъязыка, соотносимого с конкретным значением слова, используется при выборе главного варианта перевода: ПЭ, у которого признак подъязыка совпадает с тематическим индексом переводимого текста, даст наиболее близкий перевод слова к конкретной текстовой ситуации. Система меток о подъязыках, характерных для употребления слова в конкретном значении, может рассматриваться как экономное представление в словаре терминологических систем языка, поскольку вполне соответствует пониманию разработчиками АМПС понятия «термин» как употребления слова в его особом значении в текстах определенной тематической группы. При этом политерминологические слова в ПрагмП имеют несколько меток соответствующих им подъязыков (терминосистем). Многозначность термина в языке эксплицируется путем соотнесенности его с несколькими семантическими группами в других языках. Например, украинскому «відмінювання» в русском языке отвечают два ПЭ - «спряжение» и «склонение» с признаками в ПрагмП «лингв».

По грамматическим признакам «грамматический код» и «парадигматический класс» происходит интеграция в информационной лексикографической базе с грамматическими словарями соответствующих языков, используемыми в программе лемматизации и синтеза. Программа лемматизации обеспечивает возможность вхождения в словарь с текстовой словоформой. Программа синтеза парадигматических форм слова обеспечивает введение в переводной текст найденного по словарю ПЭ в форме, диктуемой правилами алгоритма МП, учитывающего структурные синтаксические соответствия языка переводимого текста и языка, на который делается перевод.

К числу других признаков, используемых при определении лексических вариантов перевода по АМПС в системе МП, относятся синтаксические признаки «тип беспредложного управления» и «тип предложного управления» - для глаголов, существительных и адъективов, признак «переходность/ непереходность» и «совершенный/ несовершенный вид» - для глагольных форм; семантические прзнаки « одушевленность», «имя собственное» - для существительных и прилагательных. Рассмотрим с этой точки зрения пример перевода по АМПС украинского многозначного слова «негідний», представленного своими лексами в двух семантических группах 6308 и 6310 со значениями соответственно «который не пригоден для чего-ниб.» и «который не заслуживает уважения». Совпадение признака «для + род.пад.» у лекса «негідний» из СГ 6308 с обнаруженной в переводимом предложении у данного слова зависимой предложно-падежной формой такого же типа будет считаться основанием для выбора в качестве главного ПЭ во всех языках семантических групп с этим же номером (6308): в английском это – unfit (синонимы improper, unsuited), в русском - неподходящий (синонимы непригодный, негодный), в немецком – ungeeignet (синонимы untauglich, unpassend, unbrauchbar, unangemessen). Совпадение признака беспредложного управления «род.пад.» у лекса «негідний» из СГ 6310 с обнаруженной в переводимом предложении у данного слова зависимой формой в родительном падеже будет считаться основанием для выбора главного ПЭ из семантических групп с номером 6310: в английском это – worthless (синонимы unworthy, despicable, base, mean, foul, vile), в русском - недостойный (синонимы низкий, подлый), в немецком – unwürdig (синонимы niederträchtig, niedrig, gemein).

Программа перевода реализует алгоритмическое определение иерархического порядка представления ЛВП с выбором среди них главного, который вводится в предложение перевода. Процедура выбора учитывает ранги лексов переводимого слова внутри семантических групп, содержащих эти лексы. Главный переводной эквивалент берется из той СГ, номер которой соответствует семантической группе, содержащей лекс переводимого слова с рангом 1.

При работе АМПС в автономном режиме, когда он выполняет функцию помощника человека-переводчика, интерфейс программы предоставляет пользователю возможность путем просмотра толкований всех семантических групп, содержащих ПЭ переводимого слова, и сопоставления их с текстовой ситуацией, в которой данное слово употребляется в исходном тексте, правильно выбрать однозначный перевод по номеру той СГ, которая по мнению переводчика наиболее адекватно описывает текстовое значение переводимого слова даже в случае, если пользователь не знает языка, на который он делает перевод. Естественно, что тогда информация о синонимах ПЭ будет невостребованной. Но эта информация может оказаться полезной при редактировании перевода человеком, знающим язык, или просто при написании пользователем текста на данном языке. Кроме того, интерфейс программы предоставляет возможность для любого знаменательного

слова из любой языковой зоны АМПС посмотреть способ образования интересующих пользователя парадигматических форм конкретного слова или всю его парадигму.

В качестве основного средства однозначного определения переводных эквивалентов в системе МП выступают двуязычные однонаправленные переводные словари словосочетаний, в которых задаются лексические и грамматические (морфолого-синтаксические) контексты конкретных значений слов исходного языка и соответствующие им ПЭ. Лексические детерминанты представлены конкретными лексемами, грамматические контексты - типовыми конструкциями с ядерным элементом, выраженным конкретной лексемой, окружение которой задается цепочками дистрибутивных грамматических классов. Употребление понятия словосочетания применительно к такому словарю в данном случае довольно условно, так как, во-первых, в Словарь наряду с действительными сочетаниями синтаксически связанных подчинительной связью слов включаются: отдельные слова (для представления ситуации, когда для переводимого слова в языке, на который оно переводится, нет однословного перевода – украинское «насправді» - русск. ПЭ «на самом деле»). Во-вторых, контекстные детерминанты могут задаваться не словами, а цепочками кодов грамматических классов: укр. мали+ГФ (инфинитив)– англ. had to + ГФ; укр. мали+ИВ (сущ., вин.пад.) – англ. had + ИВ. Данное соотношение реализуется при переводе украинских фраз «Вони мали зробити це давно» и «Вони мали час для роботи». В-третьих, в состав переводных эквивалентов для отображения дистантных связей включаются конкретные дистрибутивные классы, которые могут занимать факультативную позицию в тексте: укр. *мати*+{Наречие}+{Адъектив}+ {Адъектив}+*інформацію* – рус. *располагать*+{Наречие}+{Адъектив}+ {Адъектив}+*информацией.* Кроме того, в системе машинного перевода Словарь словосочетаний служит инструментом поиска переводных эквивалентов для определенных текстовых шаблонов, устойчивых словосочетаний, перевод которых требует либо изменения синтаксических функций переводимых единиц (укр. «при потребі» - англ. «if it is nesessary», укр. «їм слід було» - англ. «they must», укр. «визнання ухвали такою, що втратила чинність» - русск. «признание постановления утратившим силу») , либо количественного изменения ПЭ (укр. «виконати вирок» - русск. «привести приговор в исполнение".

Программа перевода по Словарю словосочетаний обеспечивает первоочередность сопоставления текстовых ситуаций с единицами словаря, репрезентирующими контактную лексическую сочетаемость, а затем проверку дистантных звязей.

Развитие АМПС будет происходить по пути интеграции с переводными словарями словосочетаний.

The Structure and Functions of the Automation Multilingual Translation Dictionary

Gryaznukhina T.A.

Questions of development of structure of the multilingual translation dictionary (MTD), its functionalities and some aspects of construction of the user interface are examined. Automatic MTD is indent for use in an independent mode, as well as for application in the systems of machine translation (MT) of scientific and technical texts. In the time of development of the dictionary structure the requirements to software of similar type were taken into account. Glossaries for each of the languages included in system of the multilingual translation embrace lexical units with their morphological, syntactic, semantic characteristics necessary for the algorithms of MT.