Воронович В.В. (Минск, БГУ)

СЛОВАРЬ ЛЕКСИЧЕСКИХ ВАЛЕНТНОСТЕЙ В СИСТЕМЕ РУССКО-БЕЛОРУССКОГО МАШИННОГО ПЕРЕВОДА

Проблема машинного перевода – одна из центральных в компьютерной лингвистике и теоретической информатике. История машинного перевода насчитывает немногим более 50 лет. За это время сменилось несколько поколений систем машинного перевода – от первых программ, использовавших ограниченные ресурсы универсальных компьютеров первого поколения до современных коммерческих продуктов, использующих мощные ресурсы серверов и персональных компьютеров. Разработка качественных и высокопроизводительных систем машинного перевода способствует разрешению важнейших социально-коммуникативных задач. Эффективность работы современной системы МП в решающей степени зависит от ее удачной настройки на конкретный подъязык естественного языка, на определенную лексику и ограниченный набор грамматических средств, характерных для текстов данной предметной области, а также на определенные типы документов.

Перевод как особый процесс межъязыковых преобразований затрагивает в комплексе разные уровни языка – морфологию, лексику, синтаксис, семантику. Модель перевода должна отражать иерархию языковых уровней, причем некоторым оптимальным для перевода образом. Опыт создания системы русско-белорусского машинного перевода в НИЛ Интеллектуальных информационных систем БГУ показал, что даже при переводе текстов на близкородственный язык необходимо анализировать текст на всех уровнях, включая семантический. Это объясняется системными несоответствиями различных аспектов языков, что приводит к невозможности использования метода прямого перевода. В частности, на грамматическом уровне можно выделить следующие несоответствия:

a) Несовпадение в категории рода имен существительных: *У нее было красивое лицо – У яе быў прыгожы твар.*
b) Несовпадение управления глаголов и существительных: *заведующий кафедрой – загадчык кафедры, гаварыць пра падзею – говорить о событии.*
c) Разные предлоги в идентичной конструкции: *по вкусу – да густу.*
d) Несоответствие синтаксических конструкций: *относящиеся к делу слова – словы, якія датычацца справы.*

Для решения всего комплекса проблем, возникающих при переводе текстов с русского языка на белорусский и с белорусского на русский, в систему машинного перевода включены следующие компоненты:

a) базовый словарь;
b) лексико-грамматический классификатор;
c) терминологические словари;

d) словарь идиом;
e) словарь сокращений;
f) словарь синтаксических валентностей глаголов;
g) совокупность правил морфологического и синтаксического анализа;
h) правила синтеза текста на языке перевода;
i) механизм настройки на предметную область.

Центральной проблемой автоматической обработки текстов на лексическом уровне до сегодняшнего дня остается проблема омонимии. Задача автоматического (реже полуавтоматического) разрешения лексической многозначности была впервые сформулирована в рамках направления науки и технологии, связанного с созданием систем машинного перевода. На сегодняшний день это критическая проблема повышения качества систем для указанного направления компьютерной лингвистики.

Различают два основных класса механизмов разрешения многозначности.

a) Механизмы автоматические, предполагающие полностью компьютерное решение этой задачи.
b) Механизмы интерактивные (диалоговые, полуавтоматические), предполагающие совместное решение задачи человеком и компьютером, и сводятся к тому, что компьютер предоставляет пользователю набор альтернатив, из которого он должен выбрать один вариант.

В создаваемой системе машинного перевода проблема разрешения омонимии решается автоматическими методами. Во многих случаях однозначный выбор эквивалента для перевода омонима производится на этапах лексико-грамматического кодирования и синтаксического анализа предложения. Однако в значительном количестве случаев омонимия остается неразрешенной даже после синтаксического анализа. Для таких случаев был создан словарь лексических валентностей, представляющий собой перечень возможной сочетаемости определенных лексем с одним из омонимов. На этапе синтаксического анализа предложения при необходимости выбора одного из нескольких вариантов перевода система проверяет, присутствуют ли возможные «спутники» омонима в данном предложении, и при наличии таковых выбирается единственно правильный эквивалент.

Каждая строка в словаре представлена искомой лексемой и ее однозначным переводом в сочетании с возможной соседней лексемой (при наличии знака «звездочка» (*) при переводе учитываются любые словоформы данной лексемы):

*рэч[=*речь] *паспаліты.

Представим систему лексической омонимии в виде трехмерного графа-классификации и покажем использование словаря валентностей в каждом комбинаторном случае. Основные признаки для омонимичной пары

– это:

a) класс словоформ (часть речи) – К,

b) грамматическая позиция словоформы – П,

c) принадлежность к одной корневой матрице – М.

Таким образом, имеется всего восемь комбинаторных вариантов при наличии или отсутствии определенных признаков:

		К П М + + +	**1**
2	К П М + + –	К П М **3** + – +	К П М **4** – + +
5	К П М + – –	К П М **6** – + –	К П М **7** – – +
		К П М – – –	**8**

Подсистема 1 в соответствии с условиями должна содержать одноматричные объекты одного класса в одной позиции. Таким условиям может отвечать только разошедшаяся полисемия. В представленном примере паре омонимов русского языка соответствуют разные эквиваленты в белорусском языке:

**положение[=*месцазнаходжанне] *планета*

**положение[=*палажэнне] *закон*.

Подсистема 2 представляет одноклассовые однопозиционные, но разнокорневые объекты:

*пшеничной муки[=*мукі]*

муки[=пакуты] совести.

Подсистема 3 содержит одноклассовые и одноматричные объекты, которые отличаются грамматической позицией (*жар**а*** и *ж**а**ра*, *кум**а*** и *к**у**ма*). Сюда будут входить также все совпадения при склонении слов неодушевлённых в именительном и винительном падежах, а у одушевлённых существительных – в родительном и винительном, например: *стаіць стол* (именительный падеж), *бачу стол* (винительный падеж). Данный вид омонимии чаще всего разрешается на этапе морфологического кодирования и синтаксического анализа.

Подсистема 4 представляет различного рода субстантивированные лексемы, некоторые из которых при переводе не являются субстантивами в переводном языке:

**следчы[=*следователь] пракуратуры*

**следчы[=*следственный] эксперымент.*

Подсистема 5 заполняется словами одной части речи, но имеющими различные корни и разные грамматические позиции:

стеклянная банка[=слоік]

национального банка [=банка]

Подсистема 6 имеет сильное ограничение в виде совпадения позиций при разных классах. Встречаются единичные случаи такого омонимичного совпадения, однако их тоже необходимо учитывать при машинном переводе:

густы[=густой] суп

мае густы [=вкусы].

Подсистема 7 представлена однокоренными лексемами разных частей речи, стоящих в различных грамматических позициях:

варта[=стоит] заўважыць

пагранічная варта[=стража]

Подсистема 8 содержит омонимы, не совпадающие ни по одному из признаков; заполнение этой подсистемы достаточно большое. Разрешение данного рода омонимии производится как на этапе синтаксического анализа, а также при помощи словаря лексических валентностей:

актавы[=актовый] зал

другой актавы[=октавы].

Таким образом, практически все комбинаторные варианты омонимии могут разрешаться при помощи словаря лексических валентностей, если с помощью грамматического кодирования и синтаксического анализа выбор однозначного эквивалента невозможен.

Литература

1. Головня А.И. Омонимия как системная категория языка: автореф. дисс. ... канд. филол. наук. Мн., 1996.
2. Зубов А.В., Зубова И.И. Основы лингвистической информатики. Часть 2. Компьютерная лингвистика, Мн., 1992.
3. Карпов В.А. Язык как система. Мн., 1992.
4. Марчук Ю.Н. Проблемы машинного перевода. М., 1983.
5. Совпель И.В. Инженерно-лингвистические принципы, методы и алгоритмы автоматической переработки текста. Мн., 1991.