Е. Б. Козеренко

ЛИНГВИСТИЧЕСКИЕ ПРЕДСТАВЛЕНИЯ В ИНТЕЛЛЕКТУАЛЬНЫХ МНОГОЯЗЫЧНЫХ СИСТЕМАХ

Текущий этап развития систем машинного перевода и систем извлечения знаний из текстов характеризуется интенсивным процессом «гибридизации» подходов и моделей, исследованиями в области когнитивной семантики, вероятностных языковых моделей и разработкой семантико-синтаксических представлений, в которых учитываются многозначность и неоднозначность синтаксических структур. Значительные вычислительные ресурсы современных систем позволяют накапливать и использовать ранее переведенные текстовые фрагменты, обеспечивать машинный перевод, основанный на прецедентах (“Example-Based Machine Translation”) [1; 2; 3], эффективно поддерживать компоненту «переводческой памяти» (“Translation

Memory") [4; 5; 6]. Создатели систем, основанных на правилах, вводят в правила веса и вероятностные характеристики, которые позволяют отобразить динамику и разнообразие языковых форм и значений, порождаемых в процессе речевой деятельности, а сторонники статистических методов построения лингвистических моделей все чаще обращаются к подходам, основанным на лингвистических знаниях, рассматривая это как средства «интеллектуализации» систем [4; 5; 6; 7; 8].

Для машинного перевода наиболее сложной проблемой является реализация языковых трансформаций, которые необходимо производить при переводе с одного языка на другой. Новое содержание проблеме языковых трансформаций придают современные реалии: необходимость проектировать и развивать обучающие компоненты систем машинного перевода и обработки текстовых знаний на основе уже существующих и вновь создаваемых корпусов параллельных текстов. При разработке компьютерных моделей русской грамматики доминировали подходы, основанные на грамматиках зависимостей [9] или локально-синтаксических представлениях, при этом не разрабатывались грамматики составляющих. Предлагаемый нами подход, разработанный в лаборатории Компьютерной лингвистики и когнитивных технологий обработки текстов ФИЦ ИУ РАН – ИПИ дает возможность компактного представления структуры составляющих предложения (грамматика фразовых структур), с одной стороны, а, с другой стороны, учитывает механизмы зависимости между узлами дерева предложения.

Одно из направлений лингвистичсекого моделирования связано с использованием функционального подхода. Разработка понятия функции, являющегося центральным в функциональной грамматике, связана с широкой проблематикой функций языка [10; 11; 12; 13; 14; 15; 16; 17; 18]. Функции связаны со значениями языковых единиц, но они не тождественны им. Исследование функции некоторой языковой формы включает анализ ее значения (или ряда значений в случае многозначности). Функциональный подход интегрирует разноуровневые языковые средства (синтаксические, лексические, словообразовательные и словоизменительные) на основе их функционально-семантических характеристик.

Отношение между понятиями категория и функция можно иначе выразить как исходное структурное значение и реализуемое структурное значение. Подобная концепция была высказана Леонардом Блумфилдом в его работе «Язык». Так, с точки зрения Л. Блумфилда, каждая лексическая форма связана с грамматическими формами в двух направлениях. С одной стороны, лексическая форма, даже когда она взята сама по себе, абстрагировано, обнаруживает значимую грамматическую структуру. С другой стороны, лексическая форма в любом конкретном высказывании, являясь особой языковой формой, всегда сопровождается той или иной грамматической формой. На основе различных функций могут возникать частично совпадающие формальные классы. Так, выполнение функции действующего лица характерно для субстантивных выражений и для типично инфинитивных словосочетаний (to scold the boys would be foolish «бранить мальчиков было бы глупо» [19].

Грамматика данного типа рассматривает в единой системе средства, относящиеся к разным языковым уровням, но объединенные на основе их семантических функций; при описании языкового материала используется подход «от семантики к ее формальному выражению» («от функций к средствам») как основной, определяющий построение грамматики, в сочетании с подходом «от формы к семантике» («от средств к функции»). Под единицами строя языка подразумеваются, прежде всего, грамматические формы слова и синтаксические конструкции, а также единицы «строевой лексики» [20]: модальные и фразовые глаголы, слова типа вчера, обычно, часто, прежде, долго и т.п.

Ю. С. Степанов [18] вводит понятие функтора как языкового средства транспозиции одного множества языковых единиц в другое множество языковых единиц того же языка. Функция есть свойство или значение функтора

Понятие функции является одним из центральных в коммуникативной грамматике Г. А. Золотовой. Функция – это предназначенность элемента к определенному способу существования в системе, к определенному служению этой системе [21]. Если за целое принимаем предложение в его коммуникативном назначении, то функции его элементов, его составных частей определяются как их строительные, комбинаторные потенции, реализуемые в построении предложения.

Функции реализуются при взаимодействии языковых объектов и их контекстов.

Для рассмотрения семантики способов конфигурирования языковых структур мы пользуемся понятием структурного знака [22], предложенным в семиотической лингвистике С. К. Шаумяном. При этом слово также рассматривается нами не с точки зрения его лексического значения, а как функционально-категориальная единица, минимальный структурный знак.

В семиотической лингвистике вводится понятие суперпозиции функций, означающее, что каждый языковой объект обладает исходной первичной функцией, а происходящие в действующем языке сдвиги значений – это наложение вторичной и других функций на исходную. Таким образом, использование инструмента суперпозиции [2; 23] категорий дает возможность выразить функциональные свойства языковых объектов.

Функциональный подход, исследующий отношения «функциональной синонимии» разнородных и разноуровневых единиц языка, чрезвычайно актуален в настоящий момент, когда проводятся эксперименты по выявлению изофункциональных и изосемичных языковых структур из параллельных текстовых корпусов [24]. Именно этот подход позволяет найти соответствия в текстах на разных языках. В самом деле, заранее нельзя с полной достоверностью определить, каким именно образом была переведена та или иная языковая структура в текстовом корпусе. Поэтому необходимо строить и исследовать различные гипотезы при проектировании лингвистического процессора.

Отсутствие полного совпадения между английскими и русскими языковыми конструкциями в научно-технических текстах можно обнаружить при изучении сравнительной частоты употребления в них отдельных частей

речи, что важно для построения систем перевода, использующих машинное обучение. Для научного изложения в целом характерен признак номинативности, т.е. более широкое использование существительных, чем в других функциональных стилях. При этом сопоставительный анализ переводов показывает, что в русском языке эта тенденция выражена более четко, и при переводе английские глаголы нередко заменяются существительными. Проведенные нами статистические исследования параллельных текстов позволяют сделать вывод о том, что русский текст приблизительно на 35 % более номинативен, чем английский. Рассмотрим следующие примеры глагольно-именных трансформаций при англо-русском переводе.

1. The fuel system is designed to store liquid gasoline and to deliver it to the engine cylinders in the form of vapor mixed with air.

Система питания предназначается для заправки жидким топливом и подачи его в цилиндры в виде смеси паров бензина с воздухом.

to store and to deliver → для заправки и подачи.

2. A similar approach has marked the EU's efforts to expand the current club of 15 countries to embrace former communist countries further east.

Точно таким же подходом характеризуются усилия ЕС по расширению нынешнего клуба 15 стран дальше на восток путем присоединения к нему бывших коммунистических стран.

to embrace→ по расширению.

Нами были проведены исследования на материале имеющихся в нашем распоряжении параллельных переводов научных статей и отдельно взятых примеров высказываний с исследуемыми конструкциями, а также мы обращались к опросу экспертов-переводчиков.

Как оказалось, в процессе грамматического разбора предложений важно учитывать вероятностные характеристики такого анализа. Рассмотрим подробнее, каким образом значения вероятности используются в процессе грамматического разбора. Например, вероятностная контекстно-свободная грамматика (PCFG – Probabilistic Context Free Grammar) и вероятностная грамматика замещения деревьев (PTSG – Probabilistic Tree Substitution Grammar) присваивают вероятность (P) каждому дереву разбора T (т.е. каждому деривату) предложения S. Эта информация является ключевой для разрешения неоднозначности синтаксических структур. Вероятность каждого возможного дерева разбора T определяется как произведение вероятностей всех правил r, используемых для развертывания каждого узла n в дереве разбора:

$$P(T,S) = \prod_{n \in T} p(r(n)) \tag{0.1}$$

Вероятность однозначного предложения (т.е. предложения, где нам не надо разрешать неоднозначность) равна вероятности единственного дерева разбора для этого предложения, т.е. P(T,S)=P(T). Вероятность же неоднозначного предложения равна сумме вероятностей всех возможных деревьев разбора ($\tau(S)$) данного предложения:

$$P(S) = \sum_{T \in \tau(S)} P(T,S) = \sum_{T \in \tau(S)} P(T) \tag{0.2}$$

Вероятность полного разбора предложения вычисляется с учетом категориальной информации для каждой головной вершины каждого узла. Пусть n – синтаксическая категория некоторого узла n, а h(n) – головная вершина узла n, m(n) – материнский узел для узла n, таким образом, мы будем вычислять вероятность p(r(n)|n, h(n)), для этого мы преобразовываем выражение (0.1) таким образом, что каждое правило становится обусловленным своей головной вершиной:

$$P(T,S) = \prod_{n \in T} p(r(n) \mid n, h(n)) \times p(h(n) \mid n, h(m(n))) \quad (0.3)$$

В нашей системе грамматики функциональные значения языковых структур определяются категориальными значениями головных вершин. Вероятностные характеристики вводятся в правила унификационной грамматики в виде весов, присваиваемых деревьям разбора. Неоднозначные и многозначные синтаксические структуры учитываются в многовариантной грамматике когнитивного трансфера (переноса). Неоднозначность является коренным свойством естественного языка и вызывает основные затруднения при создании систем машинного перевода.

Если говорить в целом о разработке лингвистичсеких процессов, то большую роль в результатах их функционирования имеет многоязычная лингвистическая база знаний.

Многоязычная лингвистическая база знаний создана и развивается в лаборатории Компьютерной лингвистики и когнитивных технологий обработки текстов ФИЦ ИУ РАН – ИПИ РАН на основе экспериментального лингвистического ресурса семантико-синтаксических представлений в лингвистических процессорах систем машинного перевода и обработки текстовых знаний. При разработке лингвистического процессора (на основе англо-русского и обратного трансфера) автором данной статьи [15] было предложено понятие полей функционального переноса (ПФП), явившихся базисом сегментации языковых структур для решения задач машинного перевода. Основная идея такого поля состоит в принятии гипотезы о том, что в основе грамматических структур лежат структуры когнитивные (ментальные фреймы); поле функционального переноса отражает взаимодействие элементов разных языковых уровней [14]. В основе многоязычной лингвистической базы знаний лежит разработка системы правил фразовых структур, отражающих также и отношения зависимости через механизм наследования атрибутов головной вершины. Как показано в [14; 15], этот подход более практичен с вычислительной точки зрения, и не применялся ранее для двуязычной ситуации. Функциональные значения языковых единиц закодированы как метки фразовых структур, и типы атрибутов-значений определяются функционально-категориальной семантикой. Множество языковых структур, представленных в виде синтактико-семантических комплексов, выстраиваются в иерархию правил. В данном случае это разновидность унификационно-порождающей грамматики, в которой структуры атрибутов и значений и их преобразования задаются в виде контекстно-свободных и мягко контекстно-зависимых продукционных правил. Отношения зависимости

реализуются через механизм головных вершин фразовых структур, а сами фразовые структуры задают линейные последовательности языковых объектов. В настоящий момент разрабатывается расширенная спецификация грамматики, в которой задаются категориально-функциональные признаки языковых объектов и структур и уточняются их вероятностные расширения.

ЛИТЕРАТУРА

1. *Brown, R. D.* Example-based machine translation in the Pangloss system / R. D. Brown // In COLING-96. – Copenhagen, 1996. – P. 169–174.
2. *Козеренко, Е. Б.* Логико-статистические методы представления языковых структур в машинном переводе / Е. Б. Козеренко // Тр. Междунар. конф. «Диалог'2005» «Уомпьютерная лингвистика и интеллектуальные технологии». – М.: Наука, 2005.
3. *Knight, K.* Automating knowledge acquisition for machine translation / K. Knight. – Al Magazine 18, 1997. – P. 81–96.
4. Compendium of Translation Software / John Hutchins (editor), Seventh (edition). August, 2003.
5. *Lagoudaki, E.* Translation Memories Survey 2006: User's Perseptions Around TM Usage / E. Lagoudaki // In Proceedings of the Translating and the Computer 28 Conference, London, 16–17 Novem. 2006. – Aslib/IMI, London, 2006. – P. 1–29.
6. *Wang, Ye-Yi.* Modelling with structures in statistical machine translation / Ye-Yi Wang, Alex Waibel. – In ACL 36/COLING 17, 1998. – P. 1357–1363.
7. *Alshawi, H.* A comparison of head transducers and transfer for a limited domain translation application / H. A Alshawi, Adam L. Buchsbaum, Fei Xia. – In ACL 35/ EACL 8, 1997. – P. 360–365.
8. *Plana, E.* SIMILIS Second-Generation Translation Memory Software / E. Plana // In Proceedings of the Translating and the Computer 28 Conference, London, 24–25 Novem. 2005. – Aslib/IMI, London, 2005.
9. *Mel'cuk, I. A.* Dependency Syntax: theory and practice / I. A. Mel'cuk. – Albany : State University of N.Y., 1988.
10. *Якобсон, Р. О.* Разработка целевой модели языка в европейской лингвистике в период между двумя войнами / Р. О. Якобсон // Новое в лингвистике. – М.. 1965. – Вып. 4. – С. 372–377.
11. *Якобсон, Р. О.* Шифтеры, глагольные категории и русский глагол / Р. О. Якобсон // Принципы типологического анализа языков различного строя. – М., 1972.
12. *Степанов, Ю. С.* имена, предикаты, предложения (семиологическая грамматика) / Ю. С. Степанов. – М. : Едиториал УРСС, 2004.
13. *Halliday, M. A. K.* An Introduction to Functional Grammar / M. A. K. Halliday; ed. by Edward Arnold. – London, 1985.
14. *Halliday M. A. K.* System and Gunction in Lanfuage / M. A. K. Halliday // Halliday M. A. K. Selected Papers. – London, 1976. – 250 p.
15. *Слюсарева, Н. А.* Проблемы функционального синтаксиса современного английского языка / Н. А. Слюсарева. – М., 1981. – 206 с.
16. *Шведова, Н. Ю.* Один из возможных путей построения функциональной грамматики русского языка / Н. Ю. Шведова // Проблемы функциональной грамматики. – М., 1985. – С. 30–37.
17. *Звегинцев, В. А.* Функция и цель в лингвистической теории / В. А. Звегинцев // Проблемы теоретической и экспериментальной лингвистики. – М., 1977. – С. 120–146.

18. *Гак, В. Г.* К типологии функциональных подходов к изучению языка / В. Г. Гак // Проблемы функциональной грамматики. – М., 1985. – С. 5–15.
19. *Блумфилд, Л.* Язык / Л. Блумфилд. – 2 изд. стереотип. – М. : Едиториал УРСС, 2002. – 608 с.
20. *Щерба, Л. В.* Языковая система и речевая деятельность / Л. В. Щерба. – Л., 1974. – 428 с.
21. *Золотова, Г. А.* Коммуникативная грамматика русского языка / Г. А. Золотова, Н. К. Онипенко, М. Ю. Сидорова. – М., 2004.
22. *Шаумян, С. К.* Семиотическая Лингвистика как Объяснительная Наука / С. К. Шаумян // Тр. Междунар. конф. «Диалог'2005» «Компьютерная лингвистика и интеллектуальные технологии». – М. : Наука, 2005. – С. 507–513.
23. *Shaumyan, S.* Sings, Mind, and Reality / S. Shaumyan. – John Benjamins Publishing Company, USA, 2006.
24. *Зубов, А. В.* Использование параллельного корпуса текстов для перевода несвободных словосочетаний / А. В. Зубов // Вестн. Московского гос. обл. ун-та. Серия Лингвистика. – 2011. – № 6. – С. 79–83.