Скромблевич В. Б.
г. Минск,
Международный университет «МИТСО»

РАЗРЕШЕНИЕ ПОЛИСЕМИИ И ОМОНИМИИ ПРИ МАШИННОМ ПЕРЕВОДЕ

Вопросы по разрешению многозначности и омонимии при машинном переводе находятся в центре внимания исследователей прикладной и компьютерной лингвистики и имеют многолетнюю историю.

Под машинным переводом в узком смысле понимается процесс перевода текста компьютером (полностью либо частично) с одного естественного языка на другой без вмешательства человека.

В широком смысле машинный перевод – это область научных исследований, находящаяся на рубеже лингвистики, математики, кибернетики, и имеющая целью построение систем, реализующих машинный перевод в узком смысле.

Эффективность автоматизированных систем машинного перевода зависит от того, в какой степени в них учитываются объективные законы функционирования языка и мышления. Выделяются следующие типы автоматизированных систем машинного перевода:

1) П-системы – системы прямого перевода. Данные системы включают лишь этапы морфологического анализа и синтеза, поэтому результат работы таких систем представляет своего рода подстрочный перевод;

2) Т-системы – системы с синтаксическим преобразованием исходного текста;

3) И-системы – системы с семантическим и прагматическим анализом. Данный тип систем считается наиболее сложным, так как включает не только лингвистическую информацию, но и экстралингвистическую информацию, т. е. семантику и прагматику предметной области [1].

Автоматический анализ текста включает ряд операций, которые компьютер выполняет над текстом на естественном языке согласно заданному алгоритму, и состоит из следующих этапов:

1) графематический анализ: выделение границ слов, предложений, абзацев и других элементов текста;

2) морфологический анализ: определение исходной формы каждого использованного в тексте слова и набора морфологических характеристик этого слова;

3) синтаксический анализ: выявление грамматической структуры предложений текста;

4) семантический анализ: определение смысла фраз. Данный этап является наиболее сложным, так как требуется установление семантических отношений между словами в тексте. «В основе семантического анализа лежит утверждение о том, что значение слова не является элементарной

семантической единицей. Оно делится на более элементарные смыслы – единицы словаря семантического языка. Эти единицы семантического языка являются своеобразными атомами, из различных комбинаций которых складываются «молекулы» – значения реальных слов естественного языка» [1, с. 49 – 50].

Одной из самых больших сложностей при обработке текстов на естественном языке является неоднозначность его единиц, проявляющаяся на всех его уровнях (морфологический, лексический, синтаксический) и выражается в явлениях полисемии, омонимии, синонимии.

Для разрешения морфологической омонимии при машинном переводе исследователи предлагают использовать такие методы, как 1) метод, основанный на правилах; 2) метод, основанный на статистике, 3) метод, основанный на машинном обучении [2, с. 70].

Суть метода, основанного на правилах, сводится к тому, что в некоторых ситуациях анализ контекста помогает понять синтаксическую структуру части предложения, а с ее помощью и формы слов. Данный метод требует ручного составления правил, для каждого из которых необходимо написать самостоятельный программный модуль. Пополнение системы правил становится все труднее с каждым новым правилом, а поэтому такие методы не получили широкого распространения.

Гораздо чаще для разрешения омонимии в современных морфологических процессорах применяются статистические методы и методы, основанные на машинном обучении.

Подсчет статистики различных вариантов слова является простейшим способом снятия морфологической неоднозначности. Например, метод простого подсчета вероятности позволяет рассчитать вероятность встретить определенную словоформу среди всех вариантов употребления в тексте [2, с. 71].

Для снятия омонимии при машинном переводе используются также и различные методы классификации. «В качестве параметров классификации могут браться грамматические параметры данного или соседних слов в некотором окне, их леммы, признаки наличия знаков препинания и проч. Выбор метода классификации во многом зависит от вкусов разработчика, для решения которой используются такие методы машинного обучения, как скрытые марковские модели, условные случайные поля, рекуррентные нейронные сети и др.» [2, с. 74].

Вышеуказанные методы применяются в таких системах морфологического анализа, как TreeTagger, Pymorghy2, MyStem.

Модуль морфологического анализа TreeTagger позиционируется как система для определения частей речи слов с возможностью настройки на любой естественный язык при наличии словаря и размеченного корпуса. Основной упор в данном процессоре сделан на разрешение морфологической омонимии и предсказание характеристик неизвестных слов. «Для снятия частеречной омонимии в TreeTagger используются решающие деревья для частей речи, обученные на размеченном корпусе. В узлах такого дерева находятся предикаты с ответом «да» или «нет» для двух предшествующих слов. При этом

в листьях хранятся значения вероятностей для возможных ответов. Для определения части речи входного слова достаточно, используя информацию о предыдущих словах, пройти по дереву от корня до листьев и выбрать наиболее вероятное значение» [2, с. 53].

В системе Pymorghy2 разрешение морфологической омонимии построено на основе корпусной статистики (если слово имеет несколько вариантов разбора, то среди всех выбирается наиболее вероятный), в то время как в системе MyStem применяются методы машинного обучения. В зависимости от входных данных MyStem снимает омонимию двумя способами: с учетом контекста и без учета контекста.

Для разрешения лексической омонимии при машинном переводе наиболее часто используется метод интерактивного разрешения неоднозначности. Принцип работы данного метода заключается в следующем. Автор текста составляет с помощью опорного толкового словаря родного (исходного) языка смысловые дополнения, а переводы слов, словосочетаний с учетом дополнений осуществляются с помощью специальных словарей исходного и целевых языков, согласованных со словарем. Значения представляются отдельными секциями, следующими за описаниями тех смысловых значений слова, для которых они являются общими, что позволяет учесть в процессе кодирования многообразие лексических и грамматических значений. Процесс смыслового кодирования исходного текста происходит в компьютере автора исходного текста с помощью служебной программы, которая содержит опорный толковый словарь исходного языка и по указаниям автора выполняет операции формирования смысловых дополнений. В процессе кодирования автор последовательно анализирует исходный текст и выделяет слово особым шрифтом в случае, если, по его мнению (или служебной программы), данное слово обладает хотя бы одним из следующих признаков:

а) данное слово является многозначным;

б) грамматическая форма данного слова и связанных с ним слов не отражает тот или иной оттенок фактического смысла текста, хотя в переводе на целевой язык данное слово и (или) связанные с ним слова могут иметь конкретные грамматические формы, выбор которых строго зависит от контекста;

в) форма глагола, причастия, деепричастия в исходном языке не отражает однозначно тот или иной характер описываемого в тексте действия и (или) состояния, достигнутого в результате действия, в то время как в том или ином целевом языке для выражения указанных оттенков действий и (или) состояний используются, в зависимости от фактического смысла текста, глаголы, причастия, деепричастия, имеющие конкретные грамматические формы;

г) данное слово вместе с некоторыми соседними словами представляет собой словосочетание, для перевода которого может потребоваться поиск среди известных словосочетаний, относящихся к данному слову, причем в некоторых случаях возможны различия в лексическом составе или в структуре, не влияющие на иносказательное значение словосочетания.

Далее служебная программа вызывает из опорного толкового словаря словарную статью, соответствующую отмеченному автором слову, затем автор

поясняет смысл этого слова, сопоставляя исходный текст с теми или иными элементами статьи. При этом программа может указывать автору на несовпадение употребления слова или словосочетания в исходном тексте и в отмеченном элементе словарной статьи, а также на пропущенные в процессе анализа исходного текста слова, которые, возможно, имеют лексическую неоднозначность [3].

Эффективность и доступность данного метода была показана и доказана научным коллективом Института проблем передачи информации РАН при составлении русско-английских и англо-русских «словарей омонимов» для системы ЭТАП-3. Идея проекта по Интерактивному разрешению лексической и синтаксической неоднозначности в системах автоматической обработки естественного языка заключалась в том, чтобы обеспечить человека, взаимодействующего с системой машинного перевода, простыми и ясными диагностическими описаниями неоднозначных лексических единиц, которые могли бы быть ему предъявлены на определенных стадиях обработки текста. Алгоритм анализа был модифицирован таким образом, чтобы любой сделанный человеком выбор исключал несовместимые варианты анализа (с возможностью возвращения к исходной позиции).

Для составления словарей омонимов выбирались слова, леммы (или некоторые словоформы) которых совпадали с леммами (словоформами) других слов, и для них были написаны диагностические комментарии и примеры. Примеры подбирались таким образом, чтобы максимально облегчить идентификацию значения слова. При этом, как отмечают авторы, подобрать для слова контексты, полностью исключающие возможность употребления его омонима/полисеманта, удается не всегда, так как контекст определяет лексическую единицу вероятностно, а не абсолютно. Поэтому в таких случаях качественные комментарии, свойственные методу интерактивного разрешения неоднозначности, приобретают особую важность. Комментарии могут включать: 1) аналитическое толкование значения слова или его существенный фрагмент; 2) маркер части речи, 3) простые синтаксические признаки, 4) синонимы и/или антонимы слова, – а также любые другие сведения о слове, его значении, синтактике или прагматике, которые могут оказаться полезны. В расчете на более продвинутых пользователей могут приводиться английские переводные эквиваленты [4, с. 61].

Авторы отмечают, что модуль разрешения лексической неоднозначности также нередко помогает справиться с синтаксической и морфологической неоднозначностями, не задавая пользователю никаких вопросов о синтаксисе или морфологии, поскольку выбор правильной лексемы удаляет многие неверные варианты синтаксической структуры.

Подводя итог, можно сделать вывод, что решение вопросов лексической, морфологический и синтаксической неоднозначностей при машинном переводе на данный момент возможно лишь частично. Процесс машинного перевода все еще предполагает разную степень активности человека в его выполнении, что обусловливает многообразие форм перевода, выбор которых зависит от его целей и условий. Улучшение качества современного машинного перевода

представляет собой трудоемкую задачу, суть которой состоит в том, чтобы сделать единицей описания отдельное лексическое значение, а технологии анализа могли бы устанавливать соответствие между исходным запросом и теми лексическими значениями, которые приемлемы для этого запроса по синтаксическим и семантическим критериям.

Список цитированных источников

1. Щипицина, Л. Ю. Информационные технологии в лингвистике : учеб. пособие / Л. Ю. Щипицина. – М. : ФЛИНТА : Наука, 2013. – 128 с.

2. Автоматическая обработка текстов на естественном языке и анализ данных : учеб. пособие / Е. И. Большакова [и др.]. – М. : Изд-во НИУ ВШЭ, 2017. – 269 с.

3. Проблемы машинного перевода [Электронный ресурс] // Российско-Таджикский ун-т. – Режим доступа: http://www.rtsu.tj/ru/faculties/filologicheskiy-fakultet/kafedry/kafedra-angliyskoyfilologii/45-04-02-napravlenie-lingvistika-programma-podgotovki-teoriya-perevoda-i-mezhkulturnoy-mezhyazykovoy/%D0%9F%D1%80%D0%BE%D0%B1%D0%BB%D0%B5%D0%BC%D1%8B%20%D0%BC%D0%B0%D1%88%D0%B8%D0%BD%D0%BD%D0%BE%D0%B3%D0%BE%20%D0%BF%D0%B5%D1%80%D0%B5%D0%B2%D0%BE%D0%B4%D0%B0.pdf. – Дата доступа: 25.05.2021.

4. Лазурский, А. В. Интерактивное разрешение лексической и синтаксической неоднозначности в системах автоматической обработки естественного языка / А. В. Лазурский [и др.] // Интернет-математика 2005. Автоматическая обработка веб-данных. – М., 2005. – С. 58–79.