www.esa-conference.ru

Методы и алгоритмы машинного перевода

Досжанова Айсулу Бауыржановна, магистрант 2 курса
факультета информационных технологий
Ниязова Розамгуль Сериковна, кандидат технических наук, доцент,
научный руководитель
Евразийского Национального университета имени Льва Николаевича Гумилёва
(Нур-Султан / Астана, Казахстан)

Ключевые слова: *машинный перевод, интернет, автоматическое понимание текстов, методы и алгоритмы, перевод, классификация.*

Машинный перевод крайне важен в наше время, в данной статье подробно раскрываются основные достижения за всё время существования машинного перевода. Рассмотрены нынешние системы машинного перевода, которые используются на практике.

Растущий интерес к статистическому подходу машинного перевода зависит от разработки эффективных алгоритмов для обучения ранее предложенных вероятностных моделей. Однако одной из открытых проблем статистического машинного перевода является разработка эффективных алгоритмов для перевода заданной входной строки.

Датой возникновения машинного перевода как научного направления выступает 1946 год. Впервые такая концепция была сформулирована директором отделения естественных наук Рокфеллерского фонда Уорреном Уивером, которая впоследствии вызвала большое количество дискуссий среди исследователей по всему миру.

Первая конференция, на которой были представлены положения о машинном переводе, состоялась в 1952 году. Исследователи из многих стран мира собрались в Массачусетском университете.

В течение двух лет после проведения конференции знания систематизировались, и лишь в 1954 году вниманию мирового сообщества была представлена первая концепция машинного перевода в Нью-Йорке.

При создании первой концепции машине подавалось около 60 предложений для перевода на русском языке, которые требовалось в короткий срок перевести на английский. Русский язык был выбран не случайно – ситуация на мировой арене предвещала начало холодной войны, поэтому требовалось ежедневно проводить анализ больших объемов информации и осуществлять переводы. В основном переводимые предложения имели узконаправленную тематику.

Уже в начале 60-х годов 20 года исследователи из Соединенных Штатов Америки активно работали над модернизацией системы машинного перевода.

Исследования продолжились в начале 70-х годов, когда появлялись первые микрокомпьютеры и было начато развитие компьютерных сетей. В результате проведенной модернизации мировому сообществу было представлено устройство для машинного перевода второго поколения. Работа машины была выстроена на основе синтаксической структуры, росло количество правил орфографии, применяемых при переводе. В процессе перевода текста машина проводила преобразование текста согласно синтактической структуре языка, на который производился перевод, после чего слова подставлялись из словаря, который был значительно расширен в модели нового поколения.

Следующий этап развития машинного перевода относится к 80-90 годам. Устройство третьего поколения осуществляло работу, основываясь не только на орфографических и морфологических правилах, но и на синтаксическом анализе, в результате чего качество перевода текста значительно возросло. При переводе зачастую возникали проблемы с семантикой текста. В дальнейшем машина была усовершенствована – на рынке появились устройства семантического типа, которые отличались наилучшим качеством переведенного текста.

На сегодняшний день все свершаемые переводы могут быть классифицированы по:

1.Жанрово-стилистическими особенностям: выделяются три функциональных разновидности перевода специальный, художественный и общественно-политический.

2.По форме перевода.

В настоящее время наибольшую распространенность и популярность получили машинные переводы, основанные на правилах и статистике.

Системы перевода, действующие на основе правил, подразделяются на две категории: интерлингвистические и трансферные.

Трансферные системы выполняют перевод в три основных этапа: производится анализ первоначального текста, дальнейший трансфер и конченый синтез. В качестве примера трансфертной системы перевода может быть названа система PROMT.

Алгоритм машинного перевода, основанного на лингвистическом анализе, сводится к восьми основным этапам. Данный алгоритм базируется на идее существования метаязыка.

На перовом этапе машина получает исходное предложение или текст, представленный в виде файла.

Следующий шаг – система разбивает полученный текст на предложения и слова. Среди наиболее сложных элементов для восприятия машиной выступают прямая речь, общепринятые и малораспространённые сокращения, имена, инициалы. Машина распознает слова, используя специальные шаблоны, которые представлены в виде буквенных и цифровых групп, а также знаков препинания и пунктуации. В процессе перевода как отдельные слова выделяются даты, сокращения. Например, слово «багрово-красный» будет производиться с учетом правил морфологического преобразования прилагательных. Данный

ЕВРАЗИЙСКОЕ НАУЧНОЕ ОБЪЕДИНЕНИЕ
EURASIAN SCIENTIFIC ASSOCIATION
www.esa-conference.ru

этап также включает в себя процедуру нормализации слов для того, чтобы осуществить их подготовку для поиска по внутреннему словарю.

На третьем этапе производится морфологический анализ, который осуществляется с учетом особенностей языка, с которого переводится текст. Каждое слово в тексте ищется в словаре, после чего ему присваиваются лексическо-грамматические характеристики, например – число, род, падеж.

Следующий шаг – проведение синтаксического анализа. Для каждого отдельного слова система проводит поиск слова, с которым оно должно быть согласовано после проведения перевода. Основное снятие многозначности осуществляется при выполнении поиска главных слов.

В дальнейшем синтаксическое дерево выстраивается при выполнении процесса распознавания лингвистических шаблонов, которые были заданы до выполнения перевода.

Процедура распознавания шаблонов проводится в несколько этапов:

1.Осуществляется проверка того или иного слова на определенную часть речи.

2.Система выясняет, не является ли проверяемое слово омонимом.

3.Производится проверка на соответствие окончаний двух зависимых слов.

4.Система получает семантические характеристики для управления предлогов и глаголов.

В случае, если система распознает какой-либо шаблон, существует несколько вариантов дальнейшей работы над элементами, которые шаблон покрывает:

1.Исключение из списка лексико-грамматических классов слов, которые не удовлетворяют заданным условиям.

2.Исключение слова из выделенного множества отдельных слов в выбранном предложении, а также дальнейшее присоединение к нему главного слова. Таким образом, два слова согласуются и становятся зависимыми друг от друга.

3.Осуществляется фиксации между словами.

На пятом этапе машина анализирует полученный текст с точки зрения семантики для разрешения многозначности слов на основе уже полученного дерева зависимостей. В начале проведения семантического анализа производится разрешение многозначности слов, являющихся базовыми. После проведения процедуры машина проведет согласование зависимых слов.

На шестом этапе машина осуществляет перевод уже построенного дерева, который включает в себя следующий перечень действий:

1.Базовые слова дерева переводятся пословно.

2.Глаголы, включенные в перечень базовых слов, обладающие изначальными характеристиками в виде рода, переводятся в совокупность глаголов одной парадигмы. Остальные глаголы не изменяются.

3.В результате выполнения перевода зависимые слова отражаются в виде совокупности различных вариантов. Лишь на этапе проведения синтеза происходит определение требуемых лексических характеристик.

На последнем этапе все части переведенного дерева зависимостей согласовываются между собой – для каждого из слов текста в словаре машина производит поиск нужной словоформы согласно классу.

Среди явных преимуществ использования таких систем – повышенная точность переведенного текста с минимальным содержанием неестественности.

Статистические системы получили широкую распространенность лишь в конце 20 века. При таком методе система обучается при помощи предоставления большого количества текстов на различных языках. При этом исходная информация одинакова по содержанию и структуре.

Так, методология системы «Яндекс.Переводчик» основывается на выполнении трех последовательных этапов:

1.Модели переводов, представляющей собой своеобразную таблицу, которая содержит информацию обо всех словах языка, на который требуется совершить перевод. При переводе текста система осуществляет учет не только отдельных слов, но и различных оборотов.

2.Модели языка, представляющей собой список наиболее встречаемых слов и оборотов, которые используются в том или ином тексте с определенной частотой.

3.Декодера, при котором выполняется поиск различных вариантов перевода того или иного текста. При этом исходная модель языка как бы «дает подсказку» декодеру, какой из вариантов перевода соответствует той или иной фразе больше всего.

Среди ключевых достоинств статистических систем для перевода – качество преобразованного текста, а также постоянное расширение перечня словарного запаса. Если в языке появляется какая-либо новая словоформа, система самостоятельно обучается и впоследствии может распознавать их при переводе.

На сегодняшний день развитие машинных систем перевода не стоит на месте: проводятся разработки с использованием современных информационных технологий, применяется корпусная лингвистика.

Литература:

1.Зубов А.В., Зубова И.И. Основы искусственного интеллекта для лингвистов. М., 2007.

2.Козеренко Е.Б. Глагольно-именные трансформации при англо-русском машинном переводе [Электронный ресурс]. Режим доступа: http://www.dialog-21.ru/digests/dialog2007/materials/html/43.htm

3.Панич Ю. В. Предварительная идентификация неоднозначного исходного текста и его перевод на другие языки с использованием системы согласованных словарей [Электронный ресурс]. Режим доступа: http://www.sciteclibrary.ru/rus/catalog/pages/9402.html

4.Семенов А.Л. Современные информационные технологии и перевод. М., 2008.