РАЗДЕЛ II. КОМПЛЕКСНОЕ ПОЗНАНИЕ СОВРЕМЕННОГО ЧЕЛОВЕКА И ОБЩЕСТВА

SECTION II. COMPLEX COGNITION OF THE MODERN PERSON AND SOCIETY

АНАЛИЗ ВОЗМОЖНОСТИ ПОВЫШЕНИЯ КАЧЕСТВА МАШИННОГО ПЕРЕВОДА НА ОСНОВЕ ТЕХНОЛОГИИ НЕЙРОННЫХ СЕТЕЙ

DOI: 10.25629/HC.2020.01.06

Гончаров В.В., Мальцева О.Л.

Военная академия Ракетных войск стратегического назначения имени Петра Великого

Москва, Россия

Аннотация. Рассматриваются основные аспекты машинного перевода, достоинства и недостатки систем машинного перевода, обосновывается возможность совершенствования качества перевода на основе технологии нейронных сетей с учетом стоимости систем и аналитической оценки погрешности перевода.

Ключевые слова: машинный перевод, системы на основе грамматических правил, статистические системы, гибридные системы, качество перевода, технология нейронных сетей, статистические методы оценки качества перевода.

С появлением 1947 году в США первых ЭВМ была высказана идея их использования для перевода текстов. Первая публичная демонстрация машинного перевода состоялась в 1954 году. Эксперимент получил широкий резонанс: начались исследования в странах по всему миру, – Англия, Болгария, ГДР, Италия, Китай, Франция, ФРГ, Япония и СССР. Пионерами теории машинного перевода в нашей считают Д.Ю. Панова, А.А. Ляпунова, И.С. Мухина. В России большой вклад в развитие машинного перевода внесла группа под руководством проф. Р.Г. Пиотровского (Российский государственный педагогический университет имени А.И. Герцена, Санкт-Петербург). За рубежом наибольшую известность приобрели труды японского ученого М. Нагао, который предложил использовать при машинном переводе уже готовые, введенные в память ЭВМ варианты прочтения исходных текстов, ранее сделанные опытными лингвистами. Больших успехов в области машинного перевода добились известные компьютерные лингвисты А. Эттингер и И. Бар-Хиллер (США), Н. Хомски (Израиль).

Машинный перевод (МП) в узком смысле – это процесс перевода некоторого текста в цифровом формате с одного естественного языка на другой с помощью специального программного обеспечения.

Машинный перевод (МП) в широком смысле – это область научных исследований, находящаяся на стыке лингвистики, математики, кибернетики и имеющая целью построение систем, реализующих машинный перевод в узком смысле.

Главной проблемой, с которой сталкиваются системы машинного перевода, является совокупная неоднозначность всех слов и всех грамматических правил. К. Буатэ сформулировал 12 проблем современного машинного перевода, которые подразделяются им на четыре класса:

концептуальные проблем,

проблемы архитектуры,

инженерные проблемы,

технические проблемы.

В основе современных систем лежит алгоритм перевода, использующий формальную грамматику языков и статистические данные. Чтобы выучить язык, система сравнивает тысячи параллельных текстов – содержащих одну и ту же информацию, но на разных языках. Для каждого изученного текста система строит список уникальных признаков. Например, редко используемые слова и специальные знаки, которые встречаются в тексте с определенной частотой.

Вместо «машинный» иногда употребляется слово автоматический, что не влияет на смысл. Однако термин автоматизированный перевод имеет совсем другое значение – при нём система просто помогает человеку переводить тексты.

Автоматизированный перевод предполагает такие формы взаимодействия как [3]:

частично автоматизированный перевод: например, использование переводчиком-человеком компьютерных словарей,

системы с разделением труда: компьютер обучен переводить только фразы жёстко заданной структуры (но делает это так, чтобы исправлять за ним не требовалось), а всё, не уложившееся в схему, отдаёт человеку.

В англоязычной терминологии также различаются термины *machine translation (англ.) MT,* – полностью автоматический перевод – *machine-aided*, и *MAT,* – автоматизированный – *machine-assisted translation (англ.)*; если же надо обозначить и то, и другое, пишут *M(A)T*.

В системах машинного перевода, как правило, три основные части: модель перевода, модель языка и декодер [4]. Модель перевода – это таблица, в которой для всех слов и фраз на одном языке перечислены возможные переводы на другой язык с указанием вероятности этих переводов. Система сравнивает не только отдельные слова, но и словосочетания из нескольких слов, идущих подряд. Модели перевода для каждой пары языков содержат миллионы пар слов и словосочетаний. Что касается модели языка, то она создается системой на этапе изучения текстов.

Переводом занимается декодер. Он проводит морфологический и синтаксический анализ текста и для каждого предложения подбирает все варианты перевода с сортировкой по убыванию вероятности. Затем все полученные варианты декодер оценивает с помощью модели языка на частоту употребления и выбирает предложение с наилучшим сочетанием вероятности и частоты.

В течение последних десятилетий неоднократно принимались попытки решить проблему неоднозначности и повысить качество результатов машинного перевода. Было доказано и опровергнуто множество теорий, что привело в итоге к появлению трех видов систем машинного перевода [1]:

системы на основе грамматических правил (Rule-Based Machine Translation, RBMT);

статистические системы (Statistical Machine Translation, SMT);

гибридные системы.

Системы на основе грамматических правил производят анализ текста, который используется в процессе перевода. Перевод производится на основе встроенных словарей для данной языковой пары, а также грамматик, охватывающих семантические, морфологические, синтаксические закономерности обоих языков. На основе всех этих данных исходный текст последовательно, предложение за предложением, преобразуется в текст на требуемом языке. Основной принцип работы таких систем – связь структур исходного и конечного текстов. Системы на основе грамматических правил часто разделяют еще на три подгруппы – системы пословного перевода, трансфертные системы и интерлингвистические системы. Преимуществами систем на основе грамматических правил являются грамматическая и синтаксическая точность, стабильность результата, возможность настройки на специфическую предметную область. К недостаткам систем на основе грамматических правил относят необходимость создания, поддержки и обновления лингвистических баз данных, трудоемкость создания такой системы, а также ее высокая стоимость.

Статистические системы при своей работе используют статистический анализ. В систему загружается текст на исходном языке и его «ручной» перевод на требуемый язык, после чего система анализирует статистику межъязыковых соответствий, синтаксических конструкций и т. д. Система является самообучаемой – при выборе варианта перевода она опирается на полученную ранее статистику. Чем больший словарь внутри языковой пары, и чем точнее он составлен, тем лучше результат статистического машинного перевода. С каждым новым переведенным текстом улучшается качество последующих переводов. Статистические системы отличаются быстротой настройки и легкостью добавления новых направлений перевода. Среди недостатков наиболее значительными являются наличие многочисленных грамматических ошибок и нестабильность перевода.

Гибридные системы машинного перевода сочетают в себе подходы, описанные ранее. Они позволят объединить все преимущества, которыми обладают статистические системы и системы, основанные на правилах. В настоящее время большинство систем являются гибридными – сочетая правила, статистику и технологию нейронных сетей. Они распознают грамматику языка и способны к автоматическому обучению, т.е. стали более интеллектуальными. Так с помощью лингвистического редактора можно просматривать варианты перевода, подключая различные словари, а использование ТМ-технологии (*Translation Memory*) позволяет запоминать и сохранять в базе знаний выполненные переводы [5]. Данная технология показала себя достаточно эффективной и широко используется профессиональными переводчиками. Так, например, программа переводчик четвертого поколения PROMT – 98 включает в себя:

PROMT – среда переводчика,

File Translator – приложение для пакетной обработки большого количества документов,

WebWien – браузер с синхронным переводом HTML – страниц.

В основе программ-переводчиков четвертого поколения лежит технология HTML-to-HTML, позволяющая переводить Web-страницы с полным сохранением форматирования и впоследствии двигаться по переведенным ссылкам.

В настоящее время существуют две концепции развития систем машинного перевода: модель «большого словаря со сложной структурой», которая заложена в большинстве программ-переводчиков и модель «смысл-текст», впервые сформулированная А.А.Ляпуновым, но не реализованная ни в одном проекте.

Системы машинного перевода можно использовать не только для работы с текстами, но и для перевода отдельных слов. Они содержат полноценные словари с подробными карточками слов и устойчивых выражений. Эти карточки система составляет на основе статистических данных, опираясь на правила языка. Для машинного словаря она отбирает только словарные формы слов и устойчивые выражения. Система проводит морфологический и синтаксический анализ, определяет часть речи, словарную форму слова и устанавливает границы словосочетаний. Эта информация помогает отсеивать неполные словосочетания. Чтобы избежать ошибок и опечаток, алгоритм, основанный на технологии машинного обучения, проверяет все потенциальные пары переводов и отсеивает ненадёжные.

Близкие по значению переводы группируются системой с помощью словарей синонимов. В них попадают слова, которые часто переводятся на другой язык одинаково или образуют словосочетания с одинаковыми словами. В результате машинный словарь получает всё, что ему необходимо знать о каждом слове и выражении: его словарную форму, часть речи, значения и синонимы. Некоторые системы для наглядности добавляют к переводам примеры, которые берут из параллельных текстов [2].

Вне зависимости от вида системы машинного перевода все они в своем составе «замыкаются» на поисковую систему, в основе которой лежат алгоритмы работы сложных динамических систем по командам типа «запрос – ответ». Безусловно, качество перевода зависит от тематики и стиля исходного текста, а также грамматической, синтаксической и лексической родственности языков, между которыми производится перевод. В настоящее время машинный

перевод художественных текстов практически всегда оказывается неудовлетворительного качества. Тем не менее для технических документов при наличии специализированных машинных словарей и некоторой настройке системы на особенности того или иного типа текстов возможно получение перевода приемлемого качества, который нуждается лишь в небольшой редакторской корректировке. Чем более формализован стиль исходного документа, тем большего качества перевода можно ожидать. Самых лучших результатов при использовании машинного перевода можно достичь для текстов, написанных в техническом (различные описания и руководства) и официально-деловом стиле. Таким образом формализация процесса генерирования запросов существенно влияет на качество перевода.

К сожалению, известное русское выражение «сколько людей, – столько и мнений» как нельзя лучше характеризует данную ситуацию и полностью исключает возможность формализации процесса поиска даже по базовым элементам. Исходя из этого, выходом из сложившейся ситуации является применение технологии искусственных нейронных сетей как одной из наиболее динамично развивающихся и реально используемых на практике ветвей теории искусственного интеллекта. Только поняв механизмы функционирования человеческого мозга и осознав перспективы их использования для управления современными системами появляется гипотетическая возможность «подойти» к машинному переводу со стороны человека-переводчика [6].

Таким образом, рассматривая процесс машинного перевода с позиций теории искусственного интеллекта, первоначально поисковой системе необходимо получить контент, а индексатору сгенерировать доступный для поиска индекс. Поисковый робот (закрепившееся в специальной литературе название), или «краулер», – это программа, которая автоматически проходит по всем ссылкам, найденным на странице, и выделяет их. Исходя из заранее заданного списка адресов, она осуществляет поиск новых документов, ещё не известных поисковой системе. Найденные новые страницы анализируются поисковой системой для дальнейшего индексирования. Этим занимается специальный модуль – индексатор, который предварительно разбивает страницы на части, применяя лексические и морфологические алгоритмы. Данные о веб-страницах хранятся в индексной базе. Индекс позволяет быстро находить информацию по запросам пользователей.

Программа-поисковик, в свою очередь, работает с файлами, полученными от индексатора. Когда пользователь вводит запрос в поисковую систему, она проверяет свой индекс и выдаёт список наиболее подходящих веб-страниц.

Анализ запроса начинается с определения языка, так как одно и то же слово на разных языках может обозначать разные вещи. Поэтому система обращает внимание на алфавит, регион и язык интерфейса пользователя, после чего поисковик переходит к морфологии и определяет, к какой части речи относятся написанные слова. Это позволяет находить документы, содержащие разные формы одних и тех же слов. Также поисковая система выделяет в запросе различные объекты – географические названия, имена людей и названия организаций, а чтобы учесть все возможные варианты, дополняет запрос новыми формулировками с тем же смыслом. Кроме того, поисковик автоматически исправляет ошибки или показывает результаты как по ошибочному, так и по исправленному запросам.

Большинство поисковых систем использует методы ранжирования и машинное обучение, чтобы выводить в начало списка «лучшие» результаты.

В более современных поисковых системах нейронные сети преобразуют поисковые запросы и заголовки веб-страниц в группы чисел – семантические векторы. Их можно сравнивать друг с другом и выдавать еще более точные результаты.

Существуют и поисковые алгоритмы, которые сравнивают векторы запросов и веб-страниц целиком – а не только их заголовков. Это позволяет системе понимать смысл страниц и верно отбирать их, когда люди описывают искомое своими словами. Для этого нейросеть преобразует тексты страниц в семантические векторы заранее – на этапе индексирования. А когда человек задаёт запрос, алгоритм сравнивает вектор запроса с уже известными ему векторами страниц.

Вместе с тем необходимо отметить, стремление исследователей искусственных нейронов «приблизить» к биологическим сопряжено с определенными трудностями. Для этого воспользуемся примером [6], заимствованным из исследований де Шуттера (de Schutter). Этот ученый в течение многих лет старался максимально точно и максимально достоверно (в мельчайших подробностях!) отразить в компьютерной модели все, что мы знаем о структуре и функционировании только одного нейрона – так называемой клетки Пуркинье. Модель де Шуттера базировалась на электрических элементах, которые (в соответствии с исследованиями Ходжкина и Хаксли, получивших Нобелевскую премию в 1963 г.) моделировали биоэлектрическую активность конкретных волокон (дендритов и аксона), а также клеточной мембраны тела нейрона. В работе де Штуттера с необычайной точностью была воспроизведена структура реальной клетки Пуркинье, а также были учтены результаты исследований Нейера и Сакмана (Нобелевская премия за 1991 г.), посвященных функционированию, так называемых ионных каналов. Построенная де Шуттером модель оказалась чрезвычайно сложной и дорогостоящей с вычислительной точки зрения. Достаточно сказать, что для построения этой модели было использовано:

1600 так называемых компартментов (фрагментов клетки, рассматриваемых как однородные структуры и содержащие конкретные химические соединения в строго определенных концентрациях);

8921 модель ионных каналов;

10 типов различных сложных математических описаний этих ионных каналов, зависящих от напряжения;

32000 дифференциальных уравнений;

19200 параметров, необходимых для оценивания настроек модели; точное описание морфологии клетки, реконструированной с помощью микроскопа.

Не приходится удивляться, что для моделирования полутора десятков секунд «жизни» такой нервной клетки потребовалось несколько десятков часов непрерывной работы большого компьютера. Следует признать чрезвычайную неэффективность полученных результатов моделирования. Тем не менее, из этих исследований сделаны однозначные выводы: попытка точного моделирования структуры и функционирования настоящего биологического нейрона оказалась удачной, но слишком дорогой и трудоемкой, чтобы аналогичным образом создавать практически полезные нейронные сети.

Разрешением сформулированного противоречия является построение систем машинного перевода по технологии нейронных сетей по принципу разумной достаточности, основным критерием которого является «цена – качество». Если необходимо, не зная языка, читать в подлиннике, получая при этом духовное наслаждение, произведения зарубежных авторов, сравнимых с А. Дюма, тогда в приоритете показатель «цена», расчет которого не представляет трудностей. В то же время показатель «качество», являясь вероятностной величиной, не всегда может отражать истинное положение дел, т.к. выбор рационального варианта построения системы необходимо проводить по значениям, отличающимися в третьем – четвертом знаке после запятой.

В рассматриваемом случае качество подобных систем может быть оценено погрешностью перевода [7]. Тогда в процессе эксплуатации систем появляется возможность набора необходимых статистических данных, позволяющих использовать методы построения эмпирических функций плотностей распределения погрешностей, реализуемых в условиях отсутствия достоверной информации о виде закона распределения. Как правило, эти методы должны основываться на использовании некоторого непараметрического факта. Под непараметрическим фактом обычно понимают свойство выборки (или ее преобразований), которое не зависит от функционального вида плотности распределения генеральной совокупности. В зависимости от того, какой непараметрический факт используется при построении оценок, получают соответствующий непараметрический метод (гистограмма, полиграмма, ряд, разложение по весовым функциям и др.), позволяющий аналитически оценить качество переведенного текста и выработать рекомендации по ее повышению.

Литература

1. Андреева А. Д., Меньшиков И. Л., Мокрушин А. А. Обзор систем машинного перевода // Молодой ученый. 2013. №12. С. 64-66.

2. Баранов А.Н. Введение в прикладную лингвистику. М.: Эдиториал УРСС 2001, 360 с.

3. Герд А.С. Прикладная лингвистика. СПб.: Изд. СПб ун-та, 2005. 268 с.

4. Зубов А.В., Зубова И.И. Информационные технологии в лингвистике. М.: Изд. центр «Академия», 2004. 208 с.

5. Семенов А.Л. Современные информационные технологии и перевод. М.: Изд. «Академия», 2008. 224 с.

6. Тадеусевич Р. и др. Элементарное введение в технологию нейронных сетей с примерами программ /Перевод с польск. Рудинского И.Д. М.: Горячая линия – Телеком, 2011. 408 с.

7. Яшин А.В. Статистические методы оценки метрологических характеристик военных эталонов. Мытищи-Ярославль: Канцлер. 2016, 126 с.

Гончаров Владимир Васильевич. E-mail: v_v_goncharov@mail.ru

Мальцева Ольга Леонидовна. AuthorID: 1041372. E-mail: olmalz@yandex.ru

Дата поступления: 10.12.2019 Дата принятия к публикации 15.01.2020

ANALYSIS OF THE POSSIBILITY OF IMPROVING THE QUALITY OF MACHINE TRANSFER BASED ON THE NEURAL NETWORK TECHNOLOGY

DOI: 10.25629/HC.2020.01.06

Goncharov V.V., Maltseva O.L.

Peter the Great Military Academy of Strategic Missile Forces

Moscow, Russia

Abstract. The main aspects of machine translation, the advantages and disadvantages of machine translation systems are considered, the possibility of improving the quality of translation based on neural network technology, taking into account the cost of systems and an analytical assessment of translation error, is substantiated.

Keywords: machine translation, systems based on grammatical rules, statistical systems, hybrid systems, translation quality, neural network technology, statistical methods for evaluating the quality of translation.

Goncharov Vladimir Vasilievich. E-mail: v_v_goncharov@mail.ru

Maltseva Olga Leonidovna. AuthorID: 1041372. E-mail: olmalz@yandex.ru

Date of receipt 10.12.2019 Date of acceptance 15.01.2020