А.И. Шемаров, Е.Г. Гриневич, А.А. Шемаров (Минск, Беларусь)

РАЗВИТИЕ ИНФОРМАЦИОННЫХ СИСТЕМ ДЛЯ РЕШЕНИЯ ЗАДАЧ МАШИННОГО ПЕРЕВОДА

Создание системы для решения задач машинного перевода напрямую зависит от уровня развития информационно-коммуникационных технологий доступных на конкретном этапе развития цивилизации. Поступательное накопление

количественных показателей технических систем приводит к появлению новых подходов к решению традиционных задач. В статье рассматриваются проблемы, связанные с технологиями обработки больших объемов данных. Авторы делают анализ существующих наиболее перспективных технологий, которые открывают новые возможности для создания систем машинного перевода.

Ключевые слова: *машинный перевод, обработка больших объемов данных, искусственных интеллект, новые методы машинного перевода.*

В нашей работе, посвященной современным возможностям информационно-коммуникационных технологий и проблемам машинного перевода, рассматривается развитие возможностей, которыми обладает цивилизация, как непрерывный процесс освоения и развития новых технологий, связанных с решением текущих задач в направлении осуществления цели достижения глобальных устремлений и мечтаний человечества, к которым несомненно относятся задачи увеличения продолжительности жизни, создания «вечных» источников энергии, транспортных систем «мгновенного перемещения», создания роботов и систем искусственного интеллекта («искинов»), заменяющих человека как при решении рутинных производственных задач, так и в области творчества, которая несомненно является прерогативой человека [1, стр. 205–206].

Рассматривая проблему создания систем машинного перевода в рамках создания систем искусственного интеллекта, необходимо отметить, что такая система является промежуточным этапом достижения идеальной цели. С точки зрения увеличения степени идеальности системы в теории развития технических систем, сформулированной Г.С. Альтшуллером, система машинного перевода не может рассматриваться как конечный этап достижения поставленной цели [2]. С точки зрения достижения идеальности системы, она должна приобрести исчезающе малые размеры или исчезнуть вообще. Поэтому решение задачи, когда человек будет воспринимать иностранный язык как родной, без какой-либо или с помощью исчезающе малой внутренней системы, не обязательно технической, скорее всего будет решаться в целом не только с помощью технических систем. Вероятнее всего решение этой задачи будет достигаться на следующих этапах развития систем перевода с использованием биологических, педагогических, психологических или иных технологий. Когда люди не подозревали о существовании компьютерных технологий, они предполагали в сказках, что поставленной цели можно достичь, выпив или съев что-нибудь [3, с. 77–80], а этот способ, по своей сути, гораздо ближе к реализации идеальной системы.

Развитие систем перевода и сопровождающих их технических средств напрямую связано с развитием информационных технологий в широком их

понимании. Информационные технологии не возникают с изобретением и использованием компьютера – на данном этапе это несомненно электронные цифровые вычислительные машины, и не исчезнут после замены существующих цифровых устройств другими технологическими решениями. Надо отметить, что цифровые системы, выполняющие обработку информации на основе математических алгоритмов, по сути дела выполняют элементарные математические операции. В процессе эволюции информационных технологий они будут заменены более эффективными системами на базе новых технологических решений. Об этом свидетельствует то, что, например, биологические системы практически по любым показателям являются на порядки более эффективными. То есть, исходя из интегральных показателей при приведении их к общему основанию, человек обрабатывает информацию гораздо более эффективно, чем существующие сегодня цифровые информационные системы, которые выполняют исключительно инструментальные функции. При этом необходимо отметить, что в целом, средний человек выполняет математические вычисления с большим трудом. Например, сложно перемножить вручную два шестнадцати разрядных десятичных числа (одна операция умножения над числами формата double [4]). Ещё более сложная задача – деление вышеозначенных чисел.

Очевидно, что информационные технологии возникли в связи с появлением развитой членораздельной речи. Появилась необходимость запоминания и передачи достаточно больших объемов информации от человека к человеку. Решение данной задачи было связано с использованием вполне определенных технологий, которые эффективно применяются и в настоящее время [5]. Это касается и любого устного народного творчества. В частности, героический эпос любого народа исполняется под музыкальный аккомпанемент. Сегодня существует множество примеров эффективного использования технологии музыкального сопровождения для передачи информации. Например, лучшим учителем года в США стал педагог, переложивший математические правила на рэп, чем серьезно стимулировал афроамериканских подростков к изучению математики. Другим примером могут быть рыбаки Ньюфаундленда, которые с детства изучают «навигационную» песню для возвращения домой вне зависимости от спутниковых систем навигации.

Говоря об информационных системах машинного перевода, необходимо отметить, что процесс их становления уходит своими корнями в глубокую древность и связан с изобретением и внедрением письменности. Уже тогда возникла необходимость в переводе текстов с одного языка на другой и для этой цели создавались словари различных типов. Упоминания о первых подобиях словарей датируются XXV веком до н.э. у шумеров. Это были так

называемые глоссы: на полях рукописей выписывались значения незнакомых слов. Первый известный нам полноценный словарь, представляющий собой отдельную книгу, появился в Китае в XX веке до н.э. Называется он Erya (爾雅 [Ěryǎ]) и состоит из 2094 словарных статей. Всего в нем растолковываются 13 113 иероглифов, написанных на 19 пянях – связках из 20-30 бамбуковых планок, размером 1 см на 20–40 см. Современные наиболее полные словари китайского языка содержат толкования около 60 000 иероглифов, а образованные носители китайского языка за свою жизнь выучивают в среднем около 10 000 иероглифов [6].

Создавались рукописные словари иностранных слов и разговорники, которые являются бесценными историческими источниками для изучения в рамках различных наук. К таким разговорникам, например, относится «Ein Rusch Boeck» –анонимный русско-нижненемецкий торговый разговорник, который сохранился в последней редакции 1568 года. Был составлен неизвестным автором или группой авторов, по всей видимости, немецко-ганзейских купцов-путешественников, представителями образованного слоя Homo Hanseaticus, с целью облегчить молодым ганзейским ученикам (sprakelerers) возможность освоения русской разговорной нормы на расстоянии [7; 8]. Использование этих рукописей позволяло увеличить количество переводчиков, потребность в которых была весьма значительна в связи с развитием торговли в рамках Ганзейского Союза. Конечно, рукописные издания не могли полностью обеспечить все возрастающие потребности человечества в обеспечении переводческой деятельности необходимыми инструментальными средствами. Потребность в информационных источниках была весьма значительной, и, конечно, не столько в переводческой деятельности. Потребность в большом количестве книг привело к созданию книгопечатания.

Возможность печатать книги привела к возникновению разнообразных словарей: толковых словарей, словарей иностранных слов и т. д. Процесс создания словарей по ряду известных причин имел значительную трудоемкость и продолжительность. Так, достаточно известный Словарь Академии Российской – первый толковый словарь русского языка, содержащий 43 357 слов в шести частях – разрабатывался на протяжении 11 лет. Работа над словарем началась в 1783 году на базе лингвистических исследований Российского собрания Академии наук и Вольного российского собрания.

В дальнейшем развитие цивилизации сопровождалось экспоненциальным увеличением производства информации. Но задачи перевода не могли быть решены людьми, обладающими исключительно лингвистическими знаниями. Требовались специалисты, которые владели не только профессиональными языковыми компетенциями, но и обладали

комплексом узкопрофессиональных знаний. Возникла потребность в технологическом обеспечении перевода средствами автоматизации, так как применение печатных версий словарей было трудоемким процессом и занимало много времени. Новые возможности для решения этой проблемы появились с развитием вычислительной техники.

Освоение человечеством электрической энергии позволило создать программно-управляемый автомат – компьютер около семидесяти лет тому назад. Создание компьютера вызвало неподдельный интерес к проблемам кибернетики в обществе и привело к появлению завышенных общественных ожиданий, чем практически всегда сопровождается появление новых технологий и систем. Однако этого нельзя достичь мгновенно. Любая технология, появившись, должна пройти путь эволюционного развития шаг за шагом. Ярким примером этого явления в механический период создания автоматов является автомат Кемпелена [9].

Вычислительная техника позволила существенно изменить наши представления о возможностях машинного перевода. Первое и важное решение заключалось в создании электронных словарей, которые существенно ускорили поиск незнакомых слов при выполнении перевода с иностранного языка. Такие словари могли использоваться даже на первых персональных компьютерах. Их уникальностью была возможность пополнения и редакции тезауруса. Наличие программ, которые автоматически генерировали подстрочный пословный перевод значительно ускоряли и облегчали процесс перевода со стороны специалиста в определенной области, даже не обладающего навыками в области перевода. То есть перевод мог осуществляться по контенту. Однако это были только первые шаги при создании систем машинного перевода.

Человеку, не обладающему развитыми лингвистическими способностями, достаточно сложно изучить и одновременно использовать несколько иностранных языков. Язык в основном используется поверхностно, и это не позволяет, сколько-либо точно, передать его «тонкости» и «богатство», что, в свою очередь, не позволяет использовать аспекты языка, накапливающиеся и формирующимися нациями в течение их многовековой истории. Существует опасность потери целых пластов культуры. Это очень отчетливо прослеживается при переводе литературных произведений. Прослеживается существенная разница между литературным переводом, выполненным профессиональным переводчиком, являющимся к тому же зачастую писателем или поэтом, или переводом, выполненном на базе подстрочника текста, полученного с помощью системы машинного перевода. Очень часто такой перевод может приводить к потере контекста литературного произведения. Поэтому, чтобы предотвратить потерю возможности достаточно корректного общения на

различных языках, требуется создание хорошо работающей системы машинного перевода.

В 60-х – 80-х годах прошлого века были исследованы возможности перевода с применением математических методов, что в дальнейшем привело к созданию ряда программ для автоматизированного перевода текстов. В настоящее время существует достаточно много программных продуктов, реализующих систему машинного перевода, как локальных программ, устанавливаемых на отдельных компьютерах или локальных сетях, так и сетевых сервисов, доступных в глобальной сети Интернет в режиме онлайн. Тем не менее, качество перевода часто не устраивает потребителей. Необходимо констатировать, что чисто алгоритмические методы перевода не позволяют достигнуть приемлемого качества машинного перевода. Поэтому в настоящее время интенсивно развиваются новые методы автоматического перевода. Одним из таких методов является метод «Translation Memory». При реализации этого метода фрагменты перевода, полученного ранее, хранят в специальной базе данных для их последующего использования при осуществлении переводов текстов подобной тематики. Однако создание такой базы вызывает значительные технические трудности. Также очень сложно оценить возможность повторного использования заранее переведенного текста, а это значит, что значительная часть такой базы данных не будет использоваться повторно. Более перспективными по оценкам экспертов считаются технологии комбинированного использования системы машинного перевода в комбинации с базой данных системы «Translation Memory», сочетающей лучшие стороны алгоритмических и статистических методов осуществления перевода.

По неформальному закону Мура количественные параметры вычислительных систем удваиваются в течение периода от одного до трех лет в среднем. И методы, которые были не применимы в прошлом, становятся применимыми в настоящее время. Поэтому можно сделать вывод о возможности замены математических алгоритмов в ближайшее время, методами позволяющего достигать требуемого результата прямой выборкой заранее полученного результата из памяти. То есть обладая высокоскоростной ассоциативной базой [10] наборов достоверных переводов предложений и фрагментов текстов, выполненных экспертами переводчиками, в зависимости от контекста предшествующего предложения, можно получить существенно приемлемые фрагменты перевода текстов, которые будут требовать минимальной постобработки. По нашим данным, представленным в работе [1], была получена оценка возможности хранения 8-ми и более миллиардов предложений на стандартном накопителе на жестких дисках размером четыре терабайта [1,

с. 210]. Отсюда мы можем свободно хранить, при необходимости сотни миллиардов переведенных предложений в облачных хранилищах. Вопрос будет заключаться только в том, каким образом мы сможем получить эту базу знаний, доступ к которой может осуществляться исключительно с использованием ассоциативных запоминающих устройств. Данная технология является одним из самых перспективных направлений развития вычислительных и информационных систем будущего.

Современные информационно-коммуникационные технологии, позволяющие использовать принципы интернет вещей (англ. 'Internet of Things', 'IoT'), облачные вычисления (англ. 'Cloud computing'), большие данные (англ. 'Big data'), Мобильные технологии (англ. 'Mobile technology') позволяют приступить к созданию технологической базы уже в настоящее время.

В облаке, представляющем собой мощные инфраструктурные решения, недоступные обычно стандартной организации, может быть создан информационный ресурс государственного или интернационального значения. В хранилище, создаваемом в рамках этого ресурса с использованием технологии больших данных, создается база переведенных фраз. Мобильные технологии и Интернет вещей позволяют пользоваться ресурсом переведенных фраз большому количеству людей, которые могут выступать как в роли пользователя, так и в роли эксперта, оценивая качество перевода фразы. Фразы, не имеющие перевода или переведенные неудачно, по мнению некоторого референтского количества пользователей, могут быть направлены на анализ и последующий перевод экспертным группам (возникновение информационного инцидента), которые могут само организовываться. При этом сбор и перевод фраз может осуществляться без участия человека при использовании интеллектуализированных сервисов, включающих поисковые системы, электронные библиотеки, боты интернет вещей. Однако на практике не решена до конца еще одна проблема, которая сдерживает использование хранилищ переведенных фраз. Эта – проблема эффективного поиска нужной фразы в многомиллиардном хранилище. Тем не менее, существуют предпосылки создания устройств быстрого доступа к таким объемам информации в ближайшем будущем.

Создание подобных проектов является актуальным и возможным с использованием современной технологической базы и возможностей сообщества, которое все более консолидируется на базе широкомасштабного внедрения мобильных информационно-коммуникационных технологий.

Библиографические ссылки

1. *Шемаров А.И.* Современные возможности информационно-коммуникационных технологий и проблемы машинного перевода / А.И. Шемаров,

Е.Г. Гриневич // Этналогія: традыцыі і сучаснасць : зб. навук. арт. Мінск : РІВШ, 2016. С. 204–211.

2. *Альтшуллер Г.С.* Творчество как точная наука. М. : Советское радио, 1979.

3. *Гримм Я.* Сказки братьев Гримм. М. ; Эксмо, 2015. 848 с.

4. 754-2008 — IEEE Standard for Floating-Point Arithmetic. Revision of ANSI/IEEE Std 754-1985 // ieeexplore.ieee.org, 2008 ISBN 978-0-7381-5752-8, doi:10.1109/IEEESTD.2008.4610935.

5. *Крез Р.* Иностранный для взрослых. Как выучить новый язык в любом возрасте. / Роджер Крез, Ричард Робертс. М. : Альпина Паблишер, 2017. С. 208.

6. Как возникли и развивались словари. URL: https://habr.com/ru/company/ yandex/blog/213461/ (дата доступа: 05.04.2019).

7. *Болек А.* Разговорники XVI–XVII вв. как источники сведений о жизни древнего Пскова: URL: http://druzhkovka-news.ru/razgovorniki-xvi-xvii-vv-kak-istochniki-svedenij-o-zhizni-drevnego-pskova/4/ (дата доступа: 05.04.2019).

8. Словарь Академіи Россійской / Том 1, 2, 3, 4, 5, 6 // Санктпетербургъ : Изд-тво при Императорской Академіи Наукъ, 1789–1794.

9. Шахматный автомат: URL: https://ru.wikipedia.org/wiki/%D0%A8%D0%B0%D1%85%D0%BC%D0%B0%D1%82%D0%BD%D1%8B%D0%B9_%D0%B0%D0%B2%D1%82%D0%BE%D0%BC%D0%B0%D1%82 (дата доступа: 05.04.2019).

10. *Кохонен Т.* Ассоциативные запоминающие устройства : пер. с англ. М. : Мир, 1982. 384 с.