ОБЗОР ИНТЕЛЛЕКТУАЛЬНЫХ МЕТОДОВ МАШИННОГО ПЕРЕВОДА

Обзор

Озерова М.И.[1,] *
[1] ORCID : 0000-0001-7658-010x;
[1] Владимирский государственный университет им. А.Г. и Н. Г. Столетовых, Владимир, Российская Федерация

* Корреспондирующий автор (ozerovam[at]rambler.ru)

Аннотация

В статье рассматриваются понятия машинного перевода. Машинный перевод, будучи одной из наиболее важных областей компьютерной лингвистики, включает в себя все проблемы обработки речи на всех языковых уровнях. Среди преимуществ машинного перевода отмечают возможность обработки большого объема данных и высокой скорости перевода при общей «нейтральности» выходных текстов. Описана история развития машинного перевода. Рассмотрены существующие технологии, описана система машинного перевода, основанного на правилах. В этой связи, в данной работе была рассмотрена классификация методов машинного перевода, перечислены их достоинства и те проблемы, с которыми исследователи сталкиваются при разработке систем машинного перевода.

Ключевые слова: машинный перевод, естественные языки, история машинного перевода, машинный перевод «по правилам», статистический машинный перевод, гибридный машинный перевод, нейронные сети.

A REVIEW OF INTELLECTUAL MACHINE TRANSLATION METHODS

Review article

Ozerova M.I.[1,] *
[1] ORCID : 0000-0001-7658-010x;
[1] Vladimir State University named after V.G. and N.G. Stoletovs, Vladimir, Russian Federation

* Corresponding author (ozerovam[at]rambler.ru)

Abstract

The article reviews the notions of machine translation. Machine translation, being one of the most important areas of computer linguistics, includes all the problems of speech processing at all linguistic levels. Among the advantages of machine translation are the ability to process large amounts of data and high translation speed with an overall "neutrality" of the output texts. The history of the development of machine translation is described. Existing technologies have been reviewed, and a rule-based machine translation system has been outlined. In this regard, this work reviewed the classification of machine translation methods, listing their merits and the problems that researchers encounter in developing machine translation systems.

Keywords: machine translation, natural languages, history of machine translation, machine translation "by the rules", statistical machine translation, hybrid machine translation, neural networks.

Введение

Перевод текстов с одного естественного языка на другой рассматривается как процесс создания на другом языке некоторого текста, эквивалентного по содержанию и способам языкового выражения исходному. При этом человек-переводчик, осуществляя перевод, истолковывая и стилистически преобразуя текст, опирается на свое видение мира, проецирует текст сквозь призму своей личности, что приводит к проявлению индивидуальности в переводе, к неизбежному отклонению от текста оригинала. Использование технологии машинного перевода позволяет эффективно решить проблему субъективных трактовок, так как машина, содержащая в базе данных множество возможных вариантов, не истолковывает, а лишь передает обнаруженные текстовые соответствия [1], [10]. Машинный перевод, будучи одной из наиболее важных областей компьютерной лингвистики, включает в себя все проблемы обработки речи на всех языковых уровнях. Среди преимуществ машинного перевода отмечают возможность обработки большого объема данных и высокой скорости перевода при общей «нейтральности» выходных текстов [2, С. 117-118]. Необходимо различать системы автоматизированного перевода(computer aided software) и машинный (автоматический) перевод. Автоматизированный перевод, к классу которого относится программное обеспечение класса translation memory, – это те программные средства, которые используются человеком-переводчиком в процессе перевода для повышения производительности труда. К машинному переводу (machine translation) относят технологии, позволяющие осуществлять перевод с одного языка на другой с помощью компьютерной программы без участия человека. При использовании программ автоматизированного перевода сокращается время, затрачиваемое на перевод, и увеличивается его качество, однако основная часть работы лежит на человеке-переводчике. Средства автоматизации перевода обеспечивают более высокое качество перевода за счет единообразия терминологии и стиля, позволяют сохранить оригинальное форматирование, создавать память перевода на основе уже переведенных текстов и их оригиналов. Машинный перевод, осуществляя перевод без участия человека, требует при этом редактирования переведенного текста, так как неизбежно возникают ошибки и неточности, связанные с самой природой естественных языков: многозначностью, контекстуальностью, синонимичностью [3].

Анализ предметной области

История развития машинного перевода берет начало в 1954 году, во время первой публичной демонстрации перевода с помощью вычислительной техники. Ранние технологии не обладали возможностями решения проблем многозначности, не проводили лингвистического анализа, а сам перевод был приближен к пословному. Второе

поколение (вплоть до 1970-х годов) характеризовалось усилением роли языковых правил. В процессе перевода осуществлялось построение синтаксической структуры для каждого предложения, основанное на правилах грамматики входного языка. Затем происходило преобразование в синтаксическую структуру выходного языка, подстановка слов из словаря и синтез предложения на выходном языке. Третье поколение (до 1980-х) ознаменовалось появлением систем семантического типа, к которым относят системы машинного перевода с теорией «Смысл ^ Текст» в основе. Суть данной теории заключалась в использовании мета-языка, который позволил бы наиболее точно передавать не только форму, но и содержание языковых знаков. Использование таких уровней, как морфологический, фонологический, синтаксический и семантический, как предполагалось, должно было существенно повысить точность и качество перевода. Однако исследователи столкнулись с определенными трудностями при осуществлении перевода, основываясь на данной теории. В частности, не удалось разрешить проблему нахождения универсального для всех естественных языков смыслового представления. Немногим позже возникли интерактивные системы машинного перевода с привлечением участия человека на разных этапах перевода: пред- и постредактирование, частично автоматизированный перевод, смешанные системы. Доминирующей технологией для конца 20-го столетия стало обучение машины посредством предоставления достаточно большого количества параллельных текстов на разных языках. Такой подход в дальнейшем получил название статистического [4]. В настоящее время различают следующие технологии машинного перевода: аналитический, статистический и нейронный машинный перевод. Аналитический машинный перевод или машинный перевод, основанный на правилах (rule-based machine translation) стал исторически первой технологией машинного перевода. В качестве основы основывается на создании связей текста на исходном языке с текстом на требуемом, сохраняя при этом оригинальное значение.

Существующие технологии

Различают три типа систем машинного перевода, основанного на правилах:

● Прямые системы (Dictionary Based Machine Translation – Машинный перевод, основанный на словаре) – в их основе лежит словарный пословный перевод, т.е. слова представлены так же, как и в словаре, содержащимся в базе данных системы.

● Трансферные системы машинного перевода, основанного на правилах (Transfer Based Machine Translation) – это один из наиболее широко используемых методов машинного перевода. В отличие от прямой модели машинного перевода, трансферный метод предполагает выполнение трех этапов в переводе: анализ исходного языка с целью определения грамматической структуры, перенос (трансфер) результирующей структуры на структуру целевого языка, генерация текста.

●Интерлингвальные системы (Interlingual RBMT Systems) – при использовании данного метода текст исходного языка представляется в виде «нейтральной» структуры, находящейся вне зависимости от каких-либо естественных языков. Текст на целевом языке формируется на основе этого нейтрального варианта. Одним из преимуществ данного метода является то, что возрастает значение языка-посредника, что и позволяет увеличить количество языков перевода [5]. Среди этапов процесса перевода выделяют морфологический анализ, объединение отдельных слов в группы, синтаксический анализ и определение посредством алгоритма каждого слова как члена предложения, синтез предложений. Рассмотрим перевод предложения “A girl eats an apple” с исходного английского языка на немецкий. На первом этапе необходимо получить информацию о каждом из исходных слов: a – неопределенный артикль, girl – существительное, eats – глагол, an – неопределенный артикль, apple - существительное. Далее необходимо получить синтаксическую информацию о глаголе “to eat”: NP-eat-NP; eat – простое настоящее время, третье лицо, единственное число, активный залог. На третьем этапе происходит парсинг исходного предложения: (NP an apple) = прямое дополнение, зависимость от глагола “to eat”. Следует отметить, что возможны варианты, когда достаточно и частичного парсинга структуры предложения для создания карты структуры выходного текста. На четвертом этапе непосредственно происходит перевод английских слов на немецкий язык: a (категория: неопределенный артикль) = ein (категория: неопределенный артикль) girl (категория: существительное) = Mädchen (категория: существительное) eat (категория: глагол) = essen (категория: глагол) an (категория: неопределенный артикль) = ein (категория: неопределенный артикль) apple (категория: существительное) = Apfel (категория: существительное). На пятом этапе происходит генерация предложения на требуемом языке: изменяются словарные формы слов (применяются другое склонение, спряжение, число). Итогом становится перевод предложения на немецкий язык: A girl eats an apple => Ein Mädchen isst einen Apfel. Несмотря на такие положительные моменты, как синтаксическая и морфологическая точность, а также возможность настройки на предметную область, аналитический перевод требует постоянного поддержания и актуализации баз данных, длителен в разработке и игнорирует контекст [6]. Однако наиболее существенным преимуществом данного метода является отсутствие необходимости использования двуязычных текстов. Это позволяет создать систему перевода для таких языков, у которых нет общих текстов или какой-либо оцифрованной информации. Кроме того, созданную единожды систему машинного перевода, основанного на правилах, можно впоследствии использовать для перечислены все известные системе слова и фразы, варианты их перевода и вероятность этих переводов. Знания системы о языке представлены в виде вероятностной модели языка. Ее использование обусловлено необходимостью выбора тех или иных вариантов и связей в зависимости от контекста. Задача декодера заключается в подборе вариантов перевода для исходного текста, сочетая между собой фразы из модели перевода и сортируя их по убыванию вероятности. Далее происходит оценивание получившихся вариантов с помощью модели языка [6]. В случае использования технологии статистического перевод возможно включение дополнительной лингвистической информации для языков с богатым словоизменением. К положительным сторонам данного метода относят быструю настройку, экономию вычислительных ресурсов за счет исключения глубокого анализа текста, а также легкость, с которой системы справляются с переводом сложных и редких слов и терминов. Тем не менее у этого подхода есть и существенные недостатки, вызванные, прежде всего, особенностями естественных языков. Низкое качество перевода отмечается у языков, принадлежащим к разным языковым семьям – в таком случае необходимо использование сложных моделей типа tree-to-tree/tree-to-string (как в случае с английским и японским языками). Также на точности выбора лексических единиц сказывается и дефицит параллельных корпусов текстов.

Помимо указанного, при использовании метода статистического машинного перевода возникают также следующие проблемы:

● Статистические аномалии – часто при внесении информации о реальном мире используются имена собственные, которые в дальнейшем могут быть ошибочно использованы при переводе, к примеру, предложение “I took the train to Paris” может быть ошибочно переведено как «Я сел на поезд в Берлин» из-за обилия вариантов “train to Berlin” в тренировочном наборе.

● Порядок слов в разных языках может существенно отличаться. Несомненно, можно синтаксически классифицировать слова в предложениях и получить обобщенную модель типа SVO (подлежащее-сказуемое-дополнение), но при этом сложно учитывать положение служебных частей речи и изменение порядка слов при смене типа предложения (например, на вопросительное).

● Слова, не вошедшие в словарь, возникают ввиду недостатка данных при обучении, различиях в морфологии. Системы статистического перевода обычно хранят такие слова как отдельные символы без создания связи с другими словоформами или фразами [8]. Две указанные выше технологии развиваются и в настоящее время и сохраняют свою актуальность. Их переплетение и слияние породило отдельный метод машинного перевода, получивший название гибридный. Утверждается, что гибридный метод машинного перевода способен использовать особенности как машинного перевода, основанного на правилах, так и статистического машинного перевода.

Перечислим некоторые подходы:

● Правила, доработанные статистикой – в данном случае словарный перевод улучшается и корректируется за счет статистических правил.

●Статистика, управляемая правилами – правила используются для предварительной обработки информации, а также на этапе статистической коррекции результирующего перевода.

Вместе с этим на пике популярности находятся методы, использующие искусственные нейронные сети (neural machine translation) перевода обучаются совместно от начала до конца, чтобы максимизировать эффективность перевода [10].

Заключение

Рассматривая проблемы машинного перевода, уместно отметить и явление культурологической непереводимости отдельных единиц перевода. Предполагается, что в случае нахождения возможности идентификации таких единиц в тексте перевода, станет возможен их анализ и внесение в базы данных программ перевода. В связи с этим Дж.К. Катфорд предлагает разработать алгоритмы, базирующиеся на правилах, позволяющих обращаться к контекстуальному значению. Для целей машинного перевода эти правила могут иметь вид операционных команд для текстуального поиска элементов, маркированных в машинном словаре специальными диакритиками с предписанием вывести в каждом конкретном случае обусловленный эквивалент. Точное выполнение таких алгоритмов, по мнению исследователя, может существенно повысить качество, корректность перевода [11]. Таким образом, в данной работы была рассмотрена классификация методов машинного перевода, перечислены их достоинства и те проблемы, с которыми исследователи сталкиваются при разработке систем машинного перевода.

Конфликт интересов

Не указан.

Conflict of Interest

None declared.

Рецензия

Винокурова Т.Н., Омский государственный университет имени Ф. М. Достоевского, Омск, Российская Федерация

Review

Vinokurova T.N., Dostoevsky Omsk State University, Omsk, Russian Federation

Список литературы / References

1. Марчук Ю.Н. Модели перевода. Модели перевода: учеб. пособие для студ. учреждений высш. проф. образования / Ю. Н. Марчук. — М.: Издательский центр «Академия», 2010. — 176 с.

2. Карасев И.В. Системы машинного перевода / И.В. Карасев, Е.А. Артюшина // Успехи современного естествознания. — 2011. — № 7. — С. 117-118

3. Морозкина Е.А. Использование информационных технологий для оптимизации процесса перевода / Е.А. Морозкина, Н.Р. Шакирова // Вестник Башкирского университета. — 2012.

4. Дроздова К. А. Машинный перевод: история, классификация, методы / К. А. Дроздова // Вестник Омского Государственного Педагогического университета // Гуманитарные исследования. — 2015.

5. Koehn P. Statistical Machine Translation / P. Koehn. — Cambridge: Cambridge University Press, 2010. — P. 15.

6. Колганов Д.С. Обзор аналитической, статистической и нейронной технологий машинного перевода / Д.С. Колганов, Е.А. Данилов // Международный студенческий научный вестник. — 2018. — № 2-3.

7. Lagarda, A.-L. Statistical Post-Editing of a Rule-Based Machine Translation System / A.-L. Lagarda, V. Alabau, F. [et al.] // Casacuberta, Proceedings of NAACL HLT 2009 : Short Papers // Association for Computational Linguistics. — Boulder, Colorado. — 2009. — P. 217–220

8. Wołk K. Polish-English Speech Statistical Machine Translation Systems for the IWSLT 2013 / K. Wołk, K. Marasek // Proceedings of the 10th International Workshop on Spoken Language Translation, Heidelberg, Germany. — 2013. — P. 113–119.

9. Хорошилов А.А. Теоретические основы и методы построения систем фразеологического машинного перевода / А.А. Хорошилов. — М., 2016.

10. Kyunghyun C. On the Properties of Neural Machine Translation: Encoder–Decoder Approaches / C. Kyunghyun, B. van Merrienboer, D. Bahdanau [et al.]. — 2014.

11. Дубровина Е.В. Основные проблемы машинного перевода / Е.В. Дубровина, А.Н. Городищева // Актуальные проблемы авиации и космонавтики. — 2014.

Список литературы на английском языке / References in English

1. Marchuk YU.N. Modeli perevoda. Modeli perevoda: ucheb. posobie dlya stud. uchrezhdenij vyssh. prof. obrazovaniya [Translation models. Models of translation: studies. manual for students. institutions of higher Prof. education] / YU. N. Marchuk. — M.: Publishing Center «Academia», 2010. — 176 p. [in Russian]

2. Karasev I.V. Sistemy mashinnogo perevoda [Machine translation systems] / I.V. Karasev, E.A. Artyushina // Uspekhi sovremennogo estestvoznaniya [Successes of modern natural science]. — 2011. — № 7. — P. 117-118 [in Russian]

3. Morozkina E.A. Ispol'zovanie informacionnyh tekhnologij dlya optimizacii processa perevoda [Using information technology to optimize the translation process] / E.A. Morozkina, N.R. SHakirova // Vestnik Bashkirskogo universiteta [Bulletin of Bashkir University]. — 2012. [in Russian]

4. Drozdova K. A. Mashinnyj perevod: istoriya, klassifikaciya, metody [Machine translation: history, classification, methods] / K. A. Drozdova // Vestnik Omskogo Gosudarstvennogo Pedagogicheskogo universiteta [Bulletin of Omsk State Pedagogical University] // Gumanitarnye issledovaniya [Humanitarian studies]. — 2015. [in Russian]

5. Koehn P. Statistical Machine Translation / P. Koehn. — Cambridge: Cambridge University Press, 2010. — P. 15.

6. Kolganov D.S. Obzor analiticheskoj, statisticheskoj i nejronnoj tekhnologij mashinnogo perevoda [Overview of analytical, statistical and neural machine translation technologies] / D.S. Kolganov, E.A. Danilov // Mezhdunarodnyj studencheskij nauchnyj vestnik [International Student Scientific Bulletin]. — 2018. — No. 2-3. [in Russian]

7. Lagarda, A.-L. Statistical Post-Editing of a Rule-Based Machine Translation System / A.-L. Lagarda, V. Alabau, F. [et al.] // Casacuberta, Proceedings of NAACL HLT 2009 : Short Papers // Association for Computational Linguistics. — Boulder, Colorado. — 2009. — P. 217–220

8. Wołk K. Polish-English Speech Statistical Machine Translation Systems for the IWSLT 2013 / K. Wołk, K. Marasek // Proceedings of the 10th International Workshop on Spoken Language Translation, Heidelberg, Germany. — 2013. — P. 113–119.

9. Horoshilov A.A. Teoreticheskie osnovy i metody postroeniya sistem frazeologicheskogo mashinnogo perevoda [Theoretical foundations and methods of constructing phraseological machine translation systems] / A.A. Horoshilov. — M., 2016.[in Russian]

10. Kyunghyun C. On the Properties of Neural Machine Translation: Encoder–Decoder Approaches / C. Kyunghyun, B. van Merrienboer, D. Bahdanau [et al.]. — 2014.

11. Dubrovina E.V. Osnovnye problemy mashinnogo perevoda [The main problems of machine translation] / E.V. Dubrovina, A.N. Gorodishcheva // Aktual'nye problemy aviacii i kosmonavtiki [Actual problems of aviation and cosmonautics]. — 2014. [in Russian]