УДК 519.765

ОБЗОР ПОДХОДОВ К МАШИННОМУ ПЕРЕВОДУ

Козина А.В.[1], Черепков Е.А.[1], Белов Ю.С.[1],

[1]*Калужский филиал федерального государственного бюджетного образовательного учреждения высшего образования «Московский государственный технический университет имени Н.Э. Баумана (национальный исследовательский университет)», Калуга, e-mail: fn1-kf@mail.ru;*

Постоянно возникающая потребность в коммуникации между людьми разных языковых групп и невозможность полного удовлетворения спроса в переводе посредством людей переводчиков, обусловили необходимость автоматизации процесса перевода. В результате появилась область машинного перевода. Машинный перевод (МП) можно определить, как область вычислительной лингвистики, которая исследует использование компьютерного программного обеспечения для перевода текста или речи с одного естественного языка (исходного языка) на другой (целевой язык). Основная задача МП состоит в том, чтобы предоставить высокий уровень качества перевода и при этом свести сложность самого процесса к минимуму. С начала возникновения данной области, было разработано множество архитектур и методов к МП. В статье исследуются существующие на данный момент методы, среди которых: подходы машинного перевода на основе правил, подходы на основе данных и гибридные подходы. Описывается общая классификация этих подходов. Приводится их характеристика, особенности и схема работы. На основе обзора была построена общая схема классификации, сделаны выводы об эффективности, актуальности данных методов и их дальнейшего практического применения.

Ключевые слова: машинный перевод, машинный перевод на основе правил, машинный перевод на основе данных, гибридный машинный перевод

REVIEW APPROACHES TO MACHINE TRANSLATION

Kozina A.V.[1], Belov Yu.S.[1], Cherepkov E.A.[1]

[1]*Bauman Moscow State Technical University (Kaluga Branch) Kaluga, e-mail:fn1-kf@mail.ru*

The constant need for communication between people of different language groups and the inability to fully meet the demand for translation by means of translators made it necessary to automate the translation process. As a result, a machine translation area appeared. Machine translation (MP) can be defined as the area of computational linguistics, which explores the use of computer software for translating text or speech from one natural language (source language) to another (target language). The main task of the MP is to provide a high level of translation quality while minimizing the complexity of the process itself. Since the beginning of the field, many architectures and methods have been developed for the MP. The article examines the currently existing methods, among which are: machine-based approaches to rule-based translation, data-based approaches, and hybrid approaches. Describes the general classification of these approaches. Their characteristics, features and scheme of work are given. On the basis of the review, a general classification scheme was constructed, conclusions were drawn on the effectiveness and relevance of these methods and their further practical application.

Keywords: machine translation, Rule based machine translation, Example based Machine translation, Hybrid Machine translation

Введение. Язык является наиболее эффективным средством коммуникации между людьми. Сегодня масштаб языкового разнообразия в мире огромен и насчитывает более 7000 языков. Обмен информацией между различными языковыми группами не возможен без перевода. Исследования показывают, что для того, чтобы перевод аутентично передавал информацию, он должен обладать определенными знаниями о языках, между

которыми он осуществляется, глубоким пониманием синтаксиса и грамматических особенностей участвующих языков, а также их словарей [1]. Первоначально перевод выполнялся только силами людей-переводчиков, однако определенные ограничения (временные, коммерческие и пр.) привели к развитию машинных переводчиков.

Машинный перевод (МП) можно определить, как область вычислительной лингвистики, которая исследует использование компьютерного программного обеспечения для перевода текста или речи с одного естественного языка на другой. Актуальность разработки качественной системы МП обусловлена его возрастающей потребностью в различных областях науки и техники, таких, как информатика, искусственный интеллект, статистика, поисковые системы, социальные сети, интеллектуальный анализ данных [2]. Основная задача МП состоит в том, чтобы предоставить высокий уровень качества перевода с одного естественного языка (исходного языка) на другой (целевой язык) и при этом свести сложность самого процесса к минимуму.

История развития МП берет свое начало в 1940 году. К настоящему времени разработано множество подходов, используемых для построения подобных систем. Целью этой статьи является обобщение всех известных подходов, требований и классификации МП. Это позволит исследователям в зависимости от текущей задачи подобрать подходящую парадигму перевода.

Существующие подходы к МП. Системы МП бывают двух типов – двуязычные и многоязычные. В двуязычном переводе система включает в себя два языка (источник и цель). Если система осуществляет перевод только с исходного языка на целевой, то такой МП является однонаправленным, в противном случае – двунаправленным. Кроме того, существуют системы, которые в процессе перевода поочередно оперируют каждым словом и те, которые оперируют блоками.

Основное деление методов машинного перевода по типу включает в себя: методы перевода на основе правил, корпусные методы машинного перевод и гибридные методы, основанные как совокупность предыдущих двух типов.

МП на основе правил. Система МП на основе правил в процессе работы использует набор грамматических правил, лексикон (двуязычный или многоязычный), а также ПО для обработки этих правил [3]. Лексикон представляется в виде словаря, ПО помогает осуществлять взаимодействие между компонентами.

В данном подходе используется иерархия Хомского [4], содержащая информацию о типе используемой грамматики. Процесс перевода заключаются в анализе исходного текста и генерации текста на целевом языке в терминах грамматики соответствующего

типа. Проверка в основном строится на синтаксических, семантических и морфологических правилах. Системы, основанные на правилах, можно представить в виде пирамиды Вокуа, изображенной на рисунке 1.

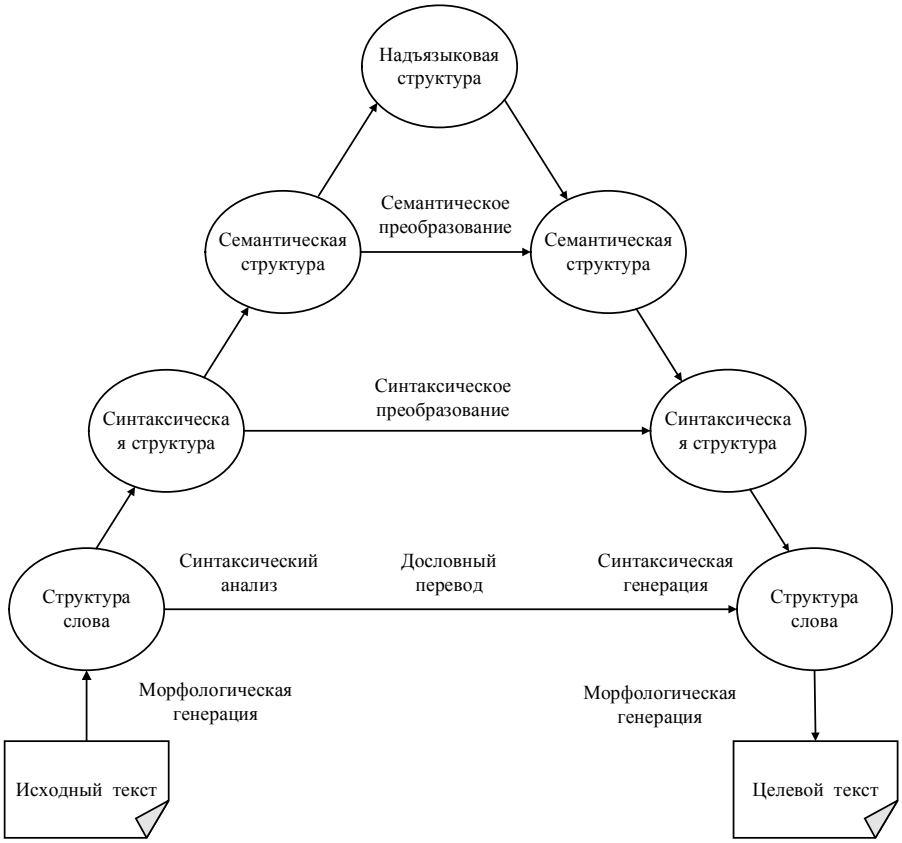

Рис. 1. Пирамида Вокуа

Использование данного подхода всегда является компромиссом между сложностью и качеством перевода. Это связано с отсутствием предела качества перевода как такового. Теоретически, каждая ошибка может быть исправлена путем реализации правила для этого конкретного случая, и поскольку нет ограничений на количество используемых правил, каждая ошибка может быть исправлена. Однако, это только в теории, поскольку количество возможных комбинаций слов или предложений и их различных значений в целом слишком велико. Поэтому при практическом применении все же следует ожидать определенного количества ошибок. Выделяют три типа перевода на основе правил, которые будут рассмотрены ниже.

Дословный МП. Данный метод относится к однонаправленным двуязычным подходам. Основная идея метода заключается в дословном, прямом переводе каждого слова. Структурный анализ исходного текста минимальный [5]. Контекст и смысл

предложения никак не учитывается. Учитываются морфологические особенности, такие как определение окончания слова или спряжения, однако довольно поверхностно.

Процесс дословного МП состоит из следующих шагов (рисунок 2) [2]:

1. Морфологический анализ исходного текста с целью определения базовых форм слова и устранения неоднозначностей.
2. Лексическое преобразование с использованием двуязычного словаря, поиск эквивалентных слов на целевом языке.
3. Незначительная корректировка порядка слов и морфологическая генерация на целевом языке.
4. Генерирование текста на целевом языке.

Рис. 2. Схема работы прямого МП

Трансферный машинный перевод. Название метода обусловлено использованием промежуточного шага – трансфера. На данном шаге происходит анализ текстов, на исходном и целевом языках с целью определения связей между словами и их возможными значениями для создания перевода более высокого качества. В данном методе имеется два промежуточных представления, одно из которых тесно связано с текстом на исходном языке (ИЯ), другое – на целевом языке (ЦЯ). Весь процесс состоит из трех этапов [5]:

1. Этап анализа. Анализ исходного текста на морфологическом, синтаксическом и семантическом уровнях с использованием словаря ИЯ. Результатом анализа является промежуточное представление текста на исходном языке.
2. Этап передачи (трансфер). Создание промежуточного представления на целевом языке на основе представления полученного на предыдущем шаге при помощи двуязычного словаря и грамматических правил.
3. Этап генерации. Создание структурной и лексической формы (семантика), корректировка форм слов (морфология), генерация предложения на целевом языке.

Процесс трансферного МП представлен на рисунке 3.

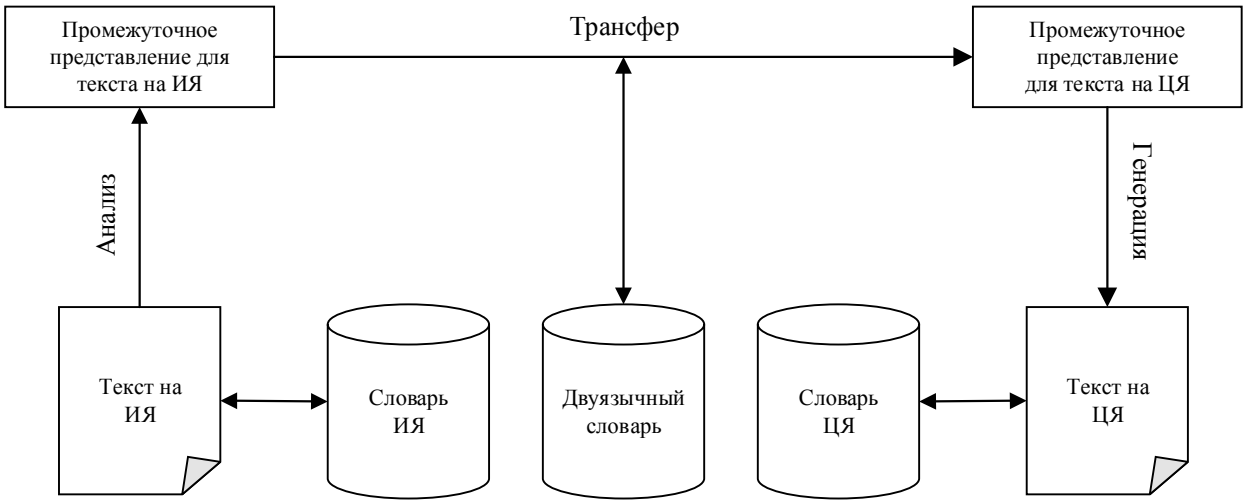

Рис. 3. Схема работы трансферного МП

Количество трансферных шагов увеличивается по мере увеличения количества языков. Для поддержки N языков, потребуется N*(N-1) таких шага.

Интерлингвальный метод МП. Основное отличие от трансферного подхода – использование вспомогательного промежуточного языка в качестве отдельного шага в процессе перевода [3]. Использование такого вспомогательного языка, который легко совместим с очень широким спектром различных естественных языков, значительно сокращает количество необходимых шагов для перевода, поскольку этапы преобразования – это преобразование из каждого исходного языка во вспомогательный язык и наоборот.

Данный подход включает в себя этапы анализа и генерации, подобные тем, что применяются в трансферном подходе. Однако, данный подход является более экономичным в количестве необходимых шагов. Поддержка N языков, потребует 2*N шага преобразования. Добавление нового языка потребует два новых этапа: этап анализа для представления на промежуточном языке и генерация предложений на целевом языке. Трансферный же перевод в этом случае потребует еще два дополнительных шага.

Процесс интерлингвального МП представлен на рисунке 4.

Рис. 4. Схема работы интерлингвального МП

Основная проблема данного подхода заключается в том, что создание такого вспомогательного языка не является тривиальной задачей, а поддержка всех существующих языков почти невозможна.

Корпусные подходы к МП. Корпусные системы МП основываются на формировании источников знаний при помощи базы данных примеров перевода и последующей автоматизации работы системы. В основе лежит двуязычная параллельная система, состоящая из согласованных корпусов данных. Основное преимущество систем данного типа заключается в том, что такие системы, при условии наличия достаточной базы примеров, достаточно быстро обучаемы. Классификация осуществляется при помощи искусственного интеллекта (ИИ), с использованием вероятности, методов статистики, кластеризации и классификации. При этом генерация такого классификатора не является «тяжелой» задачей [6]. Корпусная модель МП включает в себя статистический МП и МП на основе примеров.

Статистический МП. Статистический подход пытается отыскать шаблоны в наборе данных, рассчитать вероятности конкретного перевода и сделать разумные выводы на основе результатов. Чем выше вероятность, тем выше точность данного перевода и наоборот. Процесс статистического МП состоит из трех этапов (рисунок 5) [6]:

1. Этап языковой модели. Определение вероятности перевода на ЦЯ P(t).
2. Этап модели перевода. Определение условной вероятности выхода на ЦЯ с учетом входного текста на ИЯ P(t|s).
3. Этап декодера. Определяет наилучший перевод из возможных, путем выбора максимальных значений вероятностей, полученных на предыдущих этапах.

$$t = \arg\max\left(p\left(t \mid s\right) \cdot p\left(t\right)\right) \qquad (1)$$

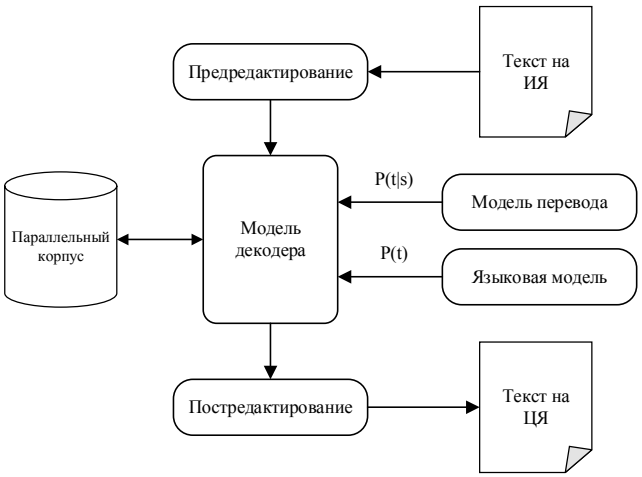

Рис. 5. Схема работы статистического МП

Статистический МП подразделяется на три вида:

1. Статистический перевод по словам. В качестве единицы перевода используются слова. Как только слова на целевом языке будут сгенерированы, происходит их переупорядочивание.
2. Статистический перевод по фразам. В качестве единицы перевода используются фразы. Из параллельного корпуса извлекаются те пары фраз, которые согласуются по выравниванию согласно принципу Филиппа Коэна, после чего оценивается вероятность полученных фраз.
3. Статистический перевод по иерархическим фразам. В основе лежит предыдущий вид перевода, но с учетом синтаксиса.

МП на основе примеров. Подход основанный на выборе данных аналогичных по смыслу и форме из базы данных примеров. Пары примеров рассматриваются в зависимости от уровня детализации: на уровне предложения, фразы, слова. В данном подходе выделяют три этапа: сопоставление, адаптация, рекомбинация [7].

1. Сопоставление. Входной текст фрагментируется в зависимости от уровня детализации. Происходит поиск примеров из базы данных близких к рассматриваемому фрагменту на ИЯ. Г. Сомерс выделяет несколько методов сопоставления: частичное соответствие, сопоставление на основе структуры, сопоставление на основе слов, на основе символов, «угол подобия» Д. Кэрролла и т.д.
2. Адаптация. Если соответствие точное, то фрагменты комбинируются для формирования вывода. Иначе извлекаются наиболее близкие найденные фрагменты, которые подвергаются выравниванию.
3. Рекомбинация. Объединение полученных фрагментов для генерации целевого текста.

Гибридный МП. МП на основе правил имеет высокую точность, но требует много ресурсов с точки зрения времени и затрат на разработку. МП на основе данных имеет высокий охват, а затраты на разработку не такие высокие. Однако для МП на основе данных потребность в корпусах является недостатком, особенно для языков с ограниченными ресурсами. Гибридный МП пытается преодолеть недостатки этих методов и одновременно с этим объединить их лучшие качества. Выделяют два подхода гибридный МП на основе данных и гибридный МП на основе правил.

Гибридный МП на основе правил. В данном подходе происходит внедрение корпусов в архитектуру МП на основе правил с целью сокращения времени разработки и стоимости. Основные концепции: использование фраз и примеров, извлеченных из

параллельного корпуса, для повышения лексики\словаря [8]; извлечение синтаксических правил и морфологии из корпуса с использованием алгоритма глубокого машинного обучения [9]; построение модуля лексического отбора с использованием преобразователя конечных состояний и модели максимальной энтропии Маркова из параллельного корпуса. Выход оценивается при помощи инструментов МП на основе данных, таких как модели языков. Наконец, внедряется система для постредактирования вывода.

Гибридный МП на основе данных. В МП на основе данных внедряется система правил. Правила применяются как к системам до\после обработки, так и к базовой системе [8]. На этапе предварительной обработки правила используются для лучшего построения целевых предложений. Морфология корректируется на этапе постредактирования при помощи алгоритмов глубокого машинного обучения [9, 10]. В рамках базовой модели системы происходит динамическая интеграция знаний синтаксиса и морфологии МП на основе правил с МП на основе данных.

Заключение. В статье был представлен обзор существующих подходов к машинному переводу, разработанных с 1940 года и их классификация. Структура классификации приведена на рисунке 6.

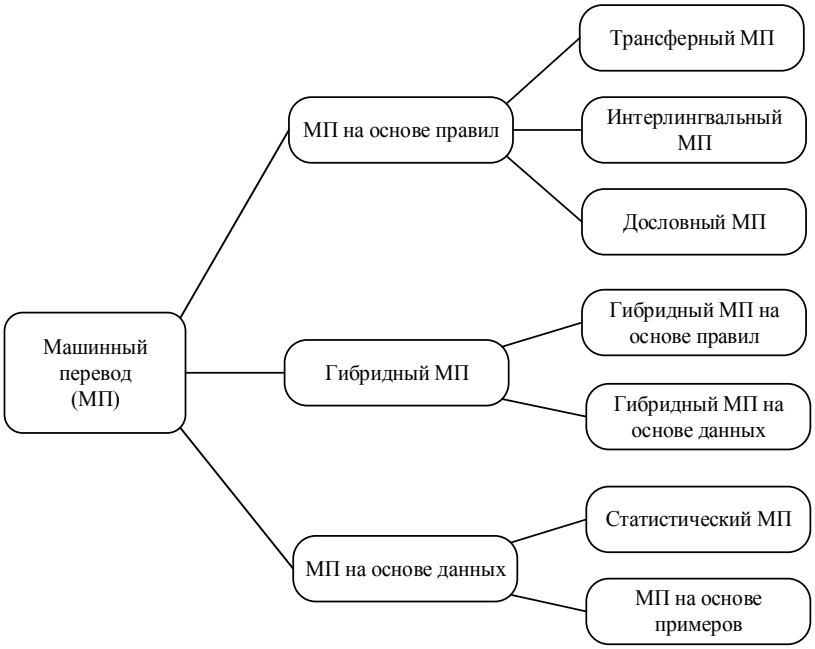

Рис. 6. Структура системы подходов к МП

Исследование показывает, что каждый из подходов обладает своими преимуществами и недостатками и в большей или меньшей степени подходит для тех или иных задач. Поскольку, язык эволюционный по своей природе; нельзя уверенно сказать, что для обработки перевода будет достаточно одного подхода. Лингвистические

нарушения, двусмысленности, отсутствие универсальности грамматики и лексики - некоторые из причин, по которым системы МП пока еще не способны предоставить превосходную точность перевода. Однако, рассмотрение существующих методов является важным для дальнейших исследований и усовершенствований в области МП, а также может заложить основу и поспособствовать в выборе того или иного подхода в иных технических и научных задачах при необходимости.

Список литературы

[1] Abiola O.B., Adetunmbi A.O., Oguntimilehin A., Using hybrid approach for English-to-Yoruba text to text machine translation system, International Journal of Computer Science and Mobile Computing, 2015, Vol. 4, No 8, pp. 308-313.

[2] Бурмистров А.В., Гришунов С.С., Молчанов А.Н. Проектирование модели системы верификации пользователя по голосу. Электронный журнал: наука, техника и образование. 2016. вып. 2. URL: http://nto-journal.ru/uploads/articles/eb6bef4d044bae56cce14c90a73482c6.pdf (дата обращения 15.11.2018)

[3] Antony P. J., Machine Translation Approaches and Survey for Indian Languages, International journal of computational Linguistics and Chinese Language Processing, 2013; Vol. 18, No. 1, pp. 47-78.

[4] Wang Y., Berwick R. C., Towards a formal framework of cognitive linguistics. In Proc. of 11th IEEE Int. Conf. Cognitive Informatics and Cognitive Computing Kyoto, Japan, 2012, pp. 1-11.

[5] Peng L., A Survey of Machine Translation Methods. TELKOMNIKA Indonesian Journal of Electrical Engineering, 2013, Vol. 11, No 12, pp. 7125-7130.

[6] Juss´a M., Farru´s M., Marin˜o J., Fonollosa J., Study and comparison of rule- based and statistical Catalan-S panish MT systems Computing and Informatics, 2012, Vol. 31, pp. 245–270.

[7] Saini Sandeep, Vineet Sahula, A Survey of Machine Translation Techniques and Systems for Indian Languages. In Computational Intelligence & Communication Technology (CICT), IEEE International Conference 2015, 2015, pp. 676-681.

[8] Costa-Jussa Marta R., Jose AR Fonollosa, Latest trends in hybrid machine translation and its applications, Computer Speech & Language, 2015, Vol. 32, No 1, pp. 3-10.

[9] Аксютина Е.М., Белов Ю.С. Обзор архитектур и методов машинного обучения для анализа больших данных. Электронный журнал: наука, техника и образование, 2016, вып.1. URL: http://nto-journal.ru/uploads/articles/0b9bd6d9833003ed0d6f9bb16fab81f1.pdf (дата обращения 28.11.2018)

[10] Рыбкин С.В., Черепков Е.А. Технологии для обработки и анализа больших данных. Электронный журнал: наука, техника и образование, 2016, вып. 4. URL:http://nto-journal.ru/uploads/articles/e87a1efcae3faba656c8818117fdfc38.pdf (дата обращения 28.11.2018)