МИНОБРНАУКИ РОССИИ **ГОСУДАРСТВЕННОЕОБРАЗОВАТЕЛЬНОЕУЧРЕЖДЕНИЕВЫСШЕГОПРОФЕССИОНАЛЬНОГООБРАЗОВАНИЯ**

**“ВОРОНЕЖСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ”**

Факультет *романо-германской филологии*

Кафедра *теории и методики преподавания немецкого языка*

*Курсовая работа*

*31. Машинный перевод*

*История развития систем машинного перевода в мире и в России*

Студент \_\_\_\_\_\_\_\_\_\_ Лежебокова М.В. 24.09.2018

Руководитель \_\_\_\_\_\_\_\_\_\_ Донина О.В

Воронеж 2018

Оглавление

[Введение 3](#_Toc526806981)

[Первые методы перевода 3](#_Toc526806982)

[Машина Троянского 3](#_Toc526806983)

[Машинный перевод на основе правил (Rule-Based Machine Translation) 5](#_Toc526806984)

[Системы дословного перевода (Direct Machine Translation) 6](#_Toc526806985)

[Трансферные системы (Transfer-based Machine Translation) 7](#_Toc526806986)

[Интерлингвистические системы (Interlingua Machine Translation) 7](#_Toc526806987)

[Машинный перевод на примерах — Example-based Machine Translation (EBMT) 7](#_Toc526806988)

[Статистический машинный перевод — Statistical Machine Translation (SMT) 9](#_Toc526806989)

[Статистический перевод по фразам — Phrase-based SMT 10](#_Toc526806990)

[Статистический перевод на основе синтаксиса — Syntax-based SMT 11](#_Toc526806991)

[Нейронный машинный перевод — Neural Machine Translation (NMT) 12](#_Toc526806992)

[Рекуррентные нейронные сети 12](#_Toc526806993)

[Кодировки 13](#_Toc526806994)

[Заключение 14](#_Toc526806995)

[Список литературы 16](#_Toc526806996)

# Введение

Машинный перевод - выполняемое на компьютере действие по преобразованию текста на одном естественном языке в эквивалентный по содержанию текст на другом языке, а также результат такого действия. Современный машинный, или автоматический перевод осуществляется с помощью человека: пред-редактора, который тем или иным образом предварительно обрабатывает подлежащий переводу текст, интер-редактора, который участвует в процессе перевода, или пост-редактора, который исправляет ошибки и недочеты в переведенном машиной тексте.

Впервые мысль о возможности машинного перевода высказал Чарльз Бэббидж (1791-1871), разработавший в 1836-1848 гг. проект цифровой аналитической машины - механического прототипа электронных цифровых вычислительных машин, появившихся через 100 лет. Идея Ч. Бэббиджа состояла в том, что память объемом 1000 50-разрядных десятичных чисел (по 50 зубчатых колес в каждом регистре) можно использовать для хранения словарей. Ч. Бэббидж привел эту идею в качестве обоснования для запроса у английского правительства средств, необходимых для физического воплощения аналитической машины, которую ему так и не удалось построить.

Для осуществления машинного перевода в компьютер вводится специальная программа, реализующая алгоритм перевода, под которым понимается последовательность однозначно и строго определенных действий над текстом для нахождения переводных соответствий в данной паре языков L1 – L2 при заданном направлении перевода (с одного конкретного языка на другой). Система машинного перевода включает в себя двуязычные словари, снабженные необходимой грамматической информацией (морфологической, синтаксической и семантической) для обеспечения передачи эквивалентных, вариантных и трансформационных переводных соответствий, а также алгоритмические средства грамматического анализа, реализующие какую-либо из принятых для автоматической переработки текста формальных грамматик. Имеются также отдельные системы машинного перевода, рассчитанные на перевод в рамках трех и более языков, но они в настоящее время являются экспериментальными.

# Первые методы перевода

## Машина Троянского

История машинного перевода начинается в 1933 году. Советский ученый Пётр Троянский обращается в Академию наук СССР с изобретённой им «машиной для подбора и печатания слов при переводе с одного языка на другой». Машина была крайне проста: большой стол, печатная машинка с лентой и плёночный фотоаппарат. На столе лежали карточки со словами и их переводами на четырёх языках.

Оператор брал первое слово из текста, находил карточку с ним, фотографировал её, а на печатной машинке набирал его морфологическую информацию: существительное, множественное число, родительный падеж. Её клавиши были модифицированы для удобства, каждая однозначно кодировала одно из свойств. Лента печатной машинки и плёнка камеры подавались параллельно, на выходе формируя набор кадров со словами и их морфологией. Полученная лента отдавалась знающим конкретные языки лингвистам, которые превращали набор фотографий в связный литературный текст. Получается, чтобы переводить тексты, как оператору, так и лингвистам требовалось знать только свой родной язык. Машина Троянского впервые на практике реализовала тот самый «промежуточный язык» (interlingua), о создании которого мечтали ещё Лейбниц и Декарт.



*Рис. 1. Машина Троянского*

В СССР изобретение признали ненужным. Никто в мире так и не знал о машине, пока его патенты не откопали в архивах двое других советских учёных в 1956 году. Произошло это неслучайно. Они искали ответ на вызов холодной войны, ведь 7 января 1954 года в штаб-квартире IBM в Нью-Йорке произошёл Джорджтаунсий эксперимент. Эта историческая находка удивила исследователей; начались изыскания. Удалось найти авторское свидетельство Троянского на «механизированный словарь», позволяющий быстро переводить тексты одновременно на несколько языков. После очередного пленарного заседания, на котором Ляпунов прочитал доклад об этом изобретении, Академией Наук был создан специальный комитет по изучению вклада Троянского. Прошло несколько лет и, наконец, в 1959 году была опубликована статья «Перeводная машина П. П. Троянского: сборник материалов о машине для перевода с одного языка на другие, предложенной П. П. Троянским в 1933 г.» за авторством И. К. Бельской и Д. Ю. Панова. Вскоре было опубликовано и авторское свидетельство, из которого было видно весьма оригинальное технологическое решение устройства.

7 января, научный компьютер IBM 701 принял участие в знаменитом Джорджтаунском эксперименте, переведя около шестидесяти русских фраз на английский. Компьютер использовал словарь из 250 слов и шесть синтаксических правил. И, конечно же, очень тщательно подобранный набор предложений, на которых проводилось тестирование.

Джорджтаунский эксперимент был одним из первых шагов в развитии машинного перевода (и одним из первых применений ЭВМ для работы с естественным языком). Тогда многие проблемы из тех, с которыми предстояло столкнуться в будущем, были еще не так очевидны. Однако главной проблемой стало то, что компьютеру труднее всего давалась работа с многозначными словами. На более-менее естественных предложениях система практически полностью переставала справляться с задачей. Сложная многокомпонентная структура таких систем также создавала проблемы: например, синтаксический анализ не всегда срабатывал верно, и составное слово guitar pick (медиатор) могло быть переведено как «выбор гитары». Также плохо переводились многозначные слова, смысл которых зависел от контекста. Например, текст «Little John was looking for his toy box. Finally he found it. The box was in the pen» вызывал (и продолжает вызывать) очень много сложностей – как словосочетание «toy box», переводившееся как «игрушечная коробка», а не «коробка для игрушек», так и «in the pen», которое переводилось как «в ручке», а не «в детском манеже». Сложности были огромными, и в итоге за 12 лет сдвинуться с мертвой точки почти не получилось. В 1966 году разгромный доклад ALPAC (Automatic Language Processing Advisory Committee) положил конец исследованиям в области машинного перевода на следующие десять лет.

## Машинный перевод на основе правил (Rule-Based Machine Translation)

Машинный перевод на основе правил (Rule-Based Machine Translation) - общий термин, который обозначает системы машинного перевода на основе лингвистической информации об исходном и переводном языках. Они состоят из двуязычных словарей и грамматик, охватывающих основные семантические, морфологические, синтаксические закономерности каждого языка. Такой подход к машинному переводу еще называют классическим. На основе этих данных исходный текст последовательно, по предложениям, преобразуется в текст перевода. Эти системы противопоставляют системам машинного перевода, которые основаны на примерах. Принцип работы таких систем — связь структуры входного и выходного предложения. Идеи машинного перевода на основе правил начали появляться ещё в 1970-х годах.

Компоненты типичной RBMT:

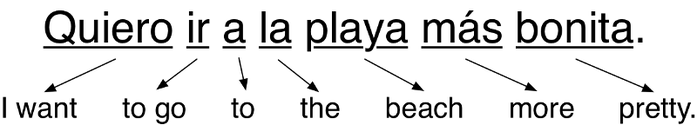
* Лингвистические базы данных: — двуязычные словари; — файлы имен, транслитерации; — морфологические таблицы.
* Модуль перевода: — грамматические правила; — алгоритмы перевода.

Особенности RBMT систем:

* Преимущества: — синтаксическая и морфологическая точность; — стабильность и предсказуемость результата; — возможность настройки на предметную область.
* Недостатки: — трудоемкость и длительность разработки, — необходимость поддерживать и актуализировать лингвистические БД; — «машинный акцент» при переводе.

### Системы дословного перевода (Direct Machine Translation)

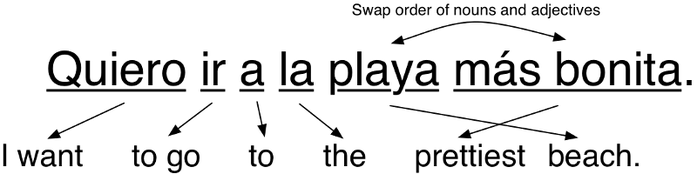
Дословный перевод: это перевод, в котором порядок слов на языке оригинала сохраняется и слова переводятся только в широком значении, не принимая во внимание контекст. Дословный перевод, при правильной передаче мысли переводимого текста, стремится к максимально близкому воспроизведению синтаксической конструкции и лексического состава подлинника. В тех случаях, когда синтаксическая структура переводимого предложения может быть и в переводе выражена аналогичными средствами, дословный перевод может рассматриваться как окончательный вариант перевода без дальнейшей литературной обработки. Однако такое совпадение синтаксических средств в двух языках встречается сравнительно редко, чаще всего при дословном переводе возникает то или иное нарушение синтаксических норм языка перевода. Несмотря на то, что дословный перевод часто нарушает синтаксические нормы русского языка, он может применяться при беглом, черновом переводе текста.



*Рис. 2. Пример дословного перевода*

### Трансферные системы (Transfer-based Machine Translation)

Трансферные системы преобразуют структуры входного языка в грамматические конструкции выходного языка. Трансферные системы включают в себя три этапа: анализ, трансфер и синтез. Для создания внутреннего представления сначала производится морфологический, лексический и семантико-синтаксический анализ входного текста. Затем для каждого предложения строится дерево разбора и производится так называемый трансфер: преобразование структуры входного предложения с учетом требований языка перевода. Последним этапом является синтез, то есть формирование выходного предложения. Классическим примером трансферной системы перевода может служить распространенная система PROMT.



*Рис. 3. Пример трансферного перевода*

### Интерлингвистические системы (Interlingua Machine Translation)

Интерлингвистический перевод - перевод в собственном смысле, т. е. передача данного содержания средствами другого естественного языка. В основе интерлингвистических систем лежит идея существования универсального метаязыка, представляющего смысл предложения на любом естественном языке. Такие системы включают в себя два этапа: анализ и синтез. На этапе анализа входной текст трансформируется при помощи словаря и грамматических правил исходного языка в представление на универсальном метаязыке. На втором этапе это представление преобразуется в предложение выходного языка при помощи словаря и грамматических правил языка перевода. Основным недостатком такого типа систем является до сих пор неразрешенная проблема нахождения универсального для всех естественных языков смыслового представления.

## Машинный перевод на примерах — Example-based Machine Translation (EBMT)

Машинный перевод на основе примеров (англ. Example-based machine translation, EBMT) — это метод машинного перевода, который часто характеризуется использованием двуязычного корпуса с параллельными текстами в качестве основной базы знаний во время выполнения перевода. По сути, это перевод по аналогии, который может рассматриваться как применение метода рассуждений на основе прецедентов к машинному обучению.

В основе машинного перевода на примерах лежит идея перевода по аналогии. Применительно к процессу перевода человеком, мысль о том, что перевод выполняется по аналогии, является отказом от идеи, что люди переводят предложения, делая глубокий лингвистический анализ. Вместо этого, данная мысль основана на убеждении, что люди переводят, сначала разбирая предложения на определённые фразы, затем переводят эти фразы, и, наконец, правильно составляют эти фрагменты в одно длинное предложение. Переводы по фразам выполняются по аналогии с предыдущими переводами. Принцип перевода по аналогии кодируется в машинном переводе на основе примеров посредством примеров переводов, которые используются для обучения такой системы. Другие подходы к машинному переводу, включая статистический машинный перевод, также используют двуязычные корпуса для изучения процесса перевода.

В общем, система EBMT состоит из трех компонентов: поиска соответствий, рекомбинации и выравнивания.

* Поиск соответствий: В соответствующем компоненте выполняется поиск множества примеров перевода для определения схожих фрагментов текстов в исходном предложении.
* Рекомбинация: На данном этапе фрагменты текста, извлечённые на этапе соответствий, объединяются для создания целого предложения. Основываясь на структуре хранения примеров (например, деревья, таблицы и т. д.), процесс объединения может потребовать конкретные процедуры для объединения текстовых единиц. Например, если примеры хранятся в древовидной структуре, для образования выходных данных следует использовать метод унификации древовидных структур.
* Выравнивание: Чтобы полностью соответствовать грамматике целевого языка и уменьшить количество несоответствий в выходных данных, необходимо выполнить некоторую последующую обработку, например, согласование подлежащего с глаголом.

Историческая важность метода была в том, что учёные всего мира впервые поняли: можно не тратить годы на создание правил и исключений, а просто взять множество уже имеющихся переводов и дать их машине. Это была ещё не революция, но шаг к ней. До революции и изобретения статистического перевода оставалось пять лет.

# Статистический машинный перевод — Statistical Machine Translation (SMT)

Когда системы, основанные на правилах, проявили свою несостоятельность, были разработаны новые подходы к переводу с использованием моделей, основанных на вероятностях и статистике, а не на грамматике.

Построение системы перевода на основе статистики требует большого количества обучающих данных, где один и тот же текст переводится как минимум на два языка. Этот дважды переведенный текст называется параллельным корпусом (parallel corpora). Точно так же, как Розеттский камень в 1800-х годах помогал ученым переводить египетские иероглифы с греческого, компьютеры могут использовать параллельные корпуса, чтобы угадать, как преобразовывать текст с одного языка в другой.

Фундаментальное отличие статистических систем перевода заключается в том, что они не пытаются создать единственно точный перевод. Вместо этого они генерируют тысячи возможных переводов, а затем ранжируют эти переводы и выбирают наиболее вероятный. Они оценивают, насколько «правильна» та или иная фраза, сравнивая ее с обучающими данными.

Сначала наше предложение разбивается на простые фрагменты, которые могут быть легко переведены. Затем мы переводим каждый из этих фрагментов, находя в обучающих данных все способы, которыми люди переводили те же самые фрагменты.

Важно отметить, что мы не просто ищем эти фрагменты в простом словаре. Мы смотрим реальные переводы реальных людей. Это помогает нам найти все возможные переводы той или иной фразы в разных контекстах:



*Рис 4. Пример статистического перевода*

Преимущества SMT:

* Быстрая настройка
* Легко добавлять новые направления перевода
* Гладкость перевода

Недостатки SMT:

* «Дефицит» параллельных корпусов
* Многочисленные грамматические ошибки
* Нестабильность перевода

## Статистический перевод по фразам — Phrase-based SMT

Взял за основу все принципы перевода по словам: статистика, перестановки и лексические упрощения. Но для обучения он разбивал текст не только на слова, но и на целые фразы. Точнее, N-граммы или фраземы — пересекающиеся наборы из N слов подряд. Машина училась переводить устойчивые сочетания слов, что заметно улучшило точность.

Хитрость метода заключалась в том, что фразы не всегда были понятными нам со школы синтаксическими конструкциями. Как только в перевод пытался вмешиваться человек, знающий про лингвистику и строение предложений, качество перевода резко падало. Пионер компьютерной лингвистики Фредерик Йелинек однажды пошутил по этому поводу: «Каждый раз, когда из команды уходит лингвист, качество распознавания возрастает».

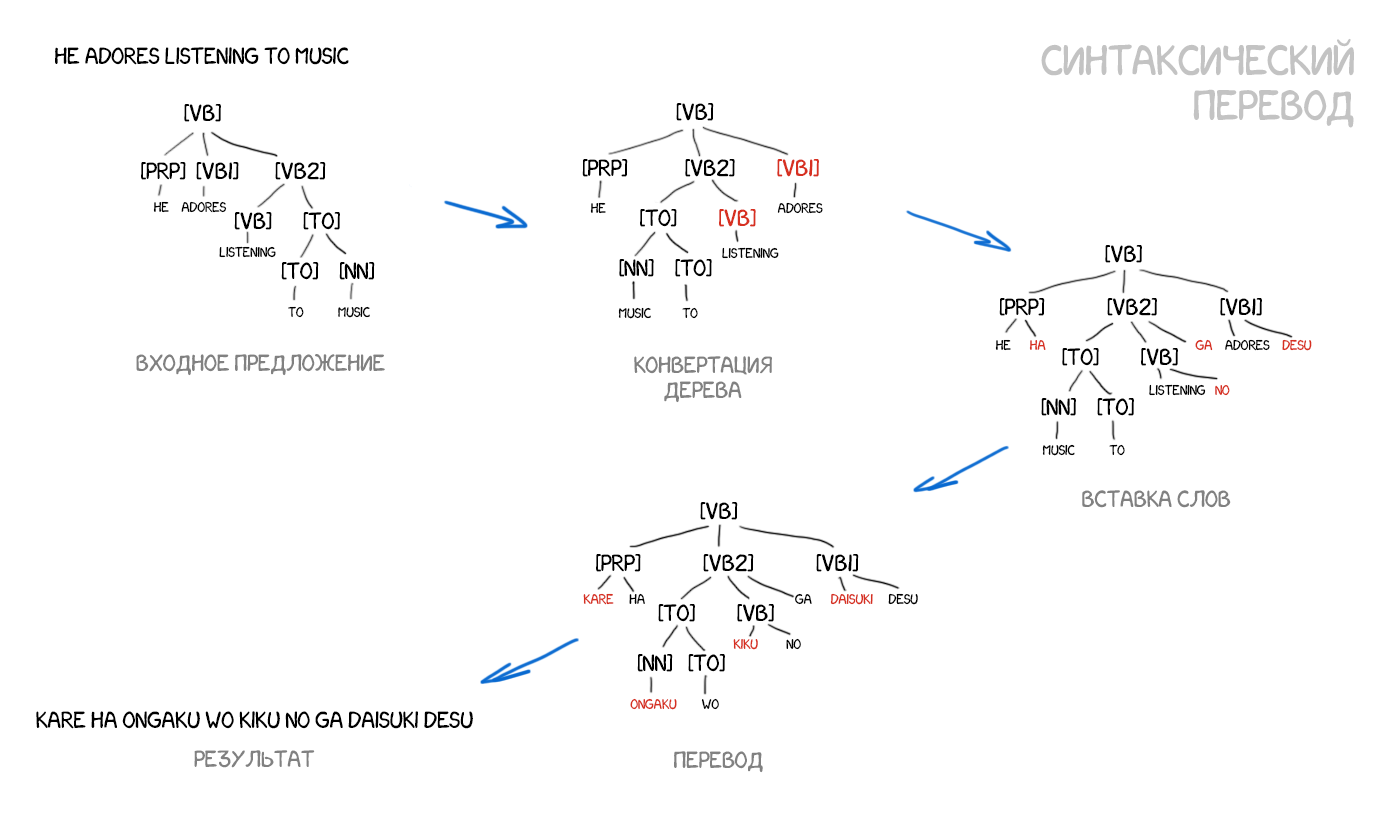
Помимо улучшения точности, перевод по фразам дал больше свободы в поиске двуязычных текстов для обучения. Для Word-based перевода было очень важно точное соответствие переводов, что исключало любые литературные или вольные переводы. Phrase-based прекрасно обучался даже на них. Многие даже начали парсить новостные сайты на разных языках и улучшать перевод текстами оттуда. С 2006 года этот подход начали использовать все. Google, «Яндекс», Bing и другие качественные онлайн-переводчики работали именно как Phrase-based до 2016-го года.

## Статистический перевод на основе синтаксиса — Syntax-based SMT

До прихода нейросетей про синтаксический перевод многие годы говорили как про «будущее переводчиков», но достичь успеха он так и не успел.

Адепты синтаксического перевода верили в объединение подходов SMT и старого трансферного перевода по правилам. Нужно научиться делать достаточно точный синтаксический разбор предложения — определять подлежащее, сказуемое, зависимые члены, а затем построить дерево.

Имея такое дерево, можно обучить машину правильно конвертировать фигуры одного языка в фигуры другого, выполняя остальной перевод по словам или фразам. Только делать это теперь не руками, а машинным обучением. В теории это решило бы проблему порядка слов навсегда. Трудность в том, что хоть человечество и считает проблему синтаксического разбора давно решённой (для многих языков есть готовые библиотеки), по факту он работает весьма плохо.



*Рис 5. Пример статистического перевода на основе синтаксиса*

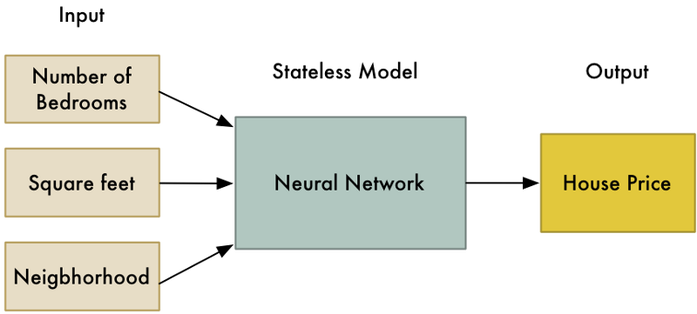
# Нейронный машинный перевод — Neural Machine Translation (NMT)

## Рекуррентные нейронные сети

В 2014 году команда KyungHyun Cho сделала прорыв. Она нашли способ применить глубокое обучение для создания системы «черного ящика». Их модель обучения принимает параллельный корпус и использует его, чтобы научиться переводить два языка без вмешательства человека.

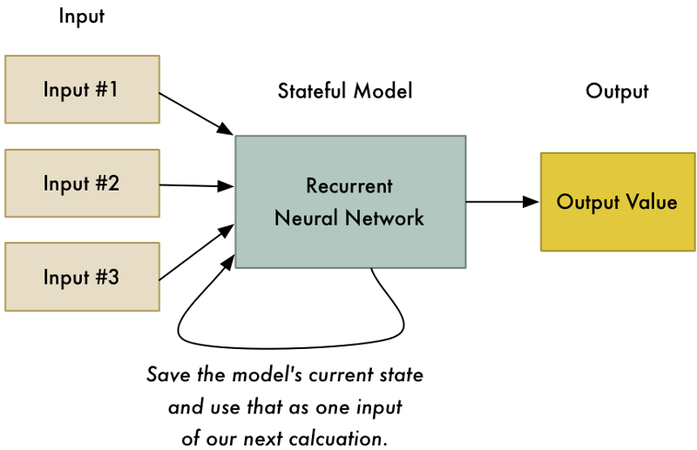
Это становится возможным благодаря двум идеям – рекуррентные нейронные сети (recurrent neural networks) и кодировки (encodings). Грамотно комбинируя эти две идеи, мы можем создать самообучающуюся систему перевода.

Обычная (нерекуррентная) нейронная сеть – это обобщенный алгоритм машинного обучения, который принимает список чисел и вычисляет результат (на основе предыдущего обучения). Нейронные сети могут использоваться в качестве «черного ящика» для решения множества проблем. Например, мы можем использовать нейронную сеть для расчета приблизительной стоимости дома на основе его параметров:



*Рис 6. Использование нейронных сетей*

Рекуррентная нейронная сеть (сокращенно RNN – Recurrent Neural Network) представляет собой немного измененную версию нейронной сети, где предыдущее состояние нейронной сети является одним из входов к следующему вычислению.



*Рис 6. Использование рекуррентных нейронных сетей*

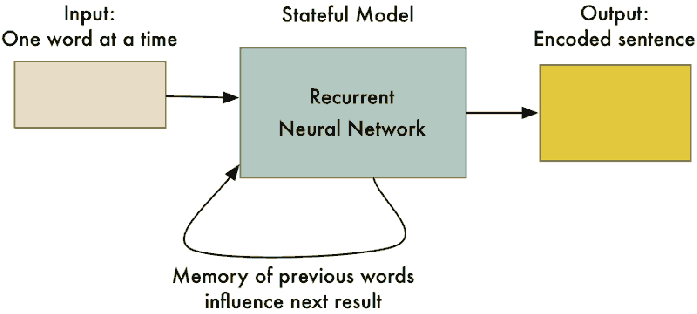
Этот трюк позволяет нейронным сетям находить закономерности и шаблоны в последовательности данных. Например, мы можем использовать его, чтобы предсказать следующее наиболее вероятное слово в предложении, если знаем первые несколько слов.

RNN полезны тогда, когда нужно работать с шаблонами данных. Поскольку человеческий язык – всего лишь один большой, сложный шаблон, RNN все чаще используются во многих областях для обработки естественного языка.

## Кодировки

Что если представить исходный текст как набор таких же характерных свойств? То есть закодировать его так, чтобы затем другая нейросеть — декодер — расшифровала его обратно в текст, но уже на другом языке. Мы специально обучим декодер знать только свой язык.

Мы можем придумать кодировку, которая представляет все возможные предложения в виде последовательности уникальных чисел. Чтобы сгенерировать кодировку, мы будем давать RNN по одному слову за раз. После обработки последнего слова мы получим значение, соответствующее всему предложению:



*Рис 7. Кодирование текста*

Мы не знаем, что означает каждый номер в кодировке, но это не имеет большого значения. Пока каждое предложение однозначно идентифицируется собственным набором чисел, нам не нужно точно знать, как эти числа были сгенерированы.

Вопрос только в том, какой вид нейросети использовать в кодере и декодере. Для картинок отлично подходят свёрточные нейросети (CNN), потому что они работают с независимыми блоками пикселей. Но в тексте не бывает независимых блоков, каждое следующее слово зависит от предыдущих и даже последующих. Текст, речь и музыка всегда последовательны.

Для их обработки лучше подходят реккурентные нейросети (RNN), ведь они помнят предыдущий результат. RNN можно применить в распознавании речи в Siri (парсинг последовательность звуков, где каждый зависит от предыдущего), подсказках слов на клавиатуре (запоминание предыдущие и угадывание следующее), генерации музыки, и так далее.

За два года нейросети превзошли всё, что было придумано в переводе за последние 20 лет. Нейронный перевод делал на 50% меньше ошибок в порядке слов, на 17% меньше лексических и на 19% грамматических ошибок. Нейросети научились сами согласовывать род и падежи в разных языках, никто их этому не учил.

# Заключение

В области машинного перевода за последние десятилетия был достигнут огромный прогресс. И, хотя до мгновенного и незаметного для пользователя перевода с любого языка галактики пока еще очень далеко, но тот факт, что за последние несколько десятилетий в этой области совершен огромный скачок, не вызывает никаких сомнений, хочется надеяться, что новые поколения систем машинного перевода будут неуклонно к нему стремиться.

# Список литературы

1. Philipp Koehn: Statistical Machine Translation.
2. Яндекс.Перевод в оффлайне. Как компьютеры научились хорошо переводить – (<https://habr.com/company/yandex/blog/224445/>)
3. Машинное обучение это весело! Часть 5 – (<https://goo.gl/KyfLnN>)
4. Understanding LSTM Networks – (<https://colah.github.io/posts/2015-08-Understanding-LSTMs/>)