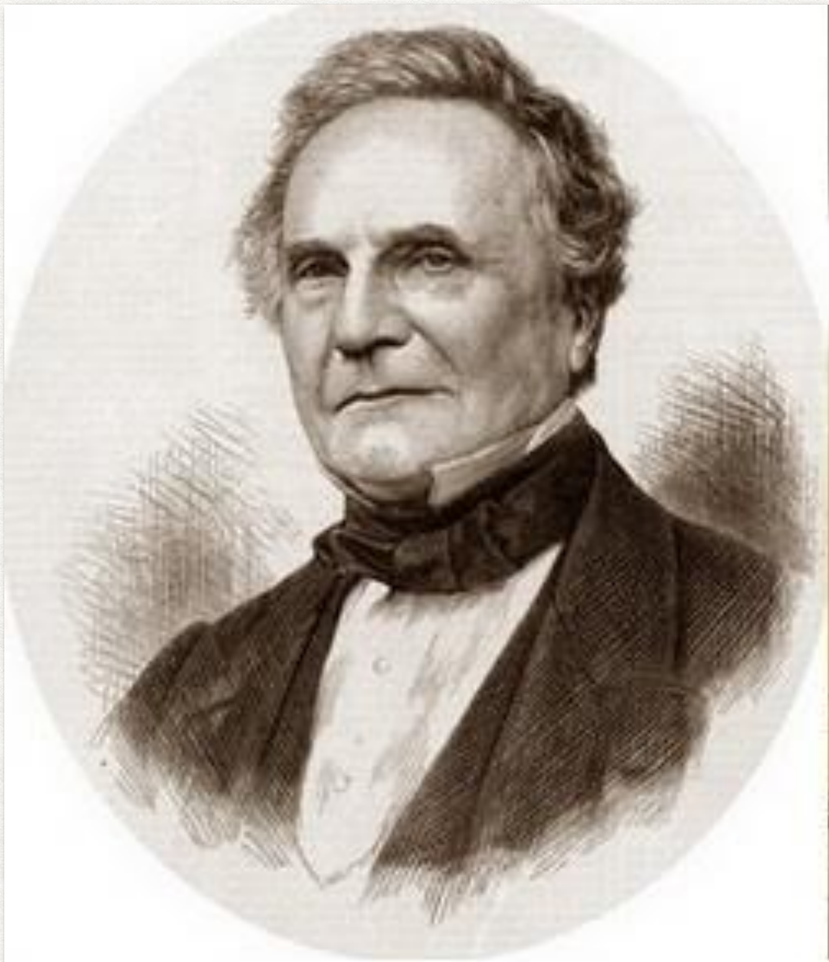
МАШИННЫЙ ПЕРЕВОД

КРАТКАЯ ИСТОРИЯ МАШИННОГО ПЕРЕВОДА

ЧАРЛЬЗ БЭББИДЖ

1836-1848 ГГ.

- Проект цифровой аналитической машины



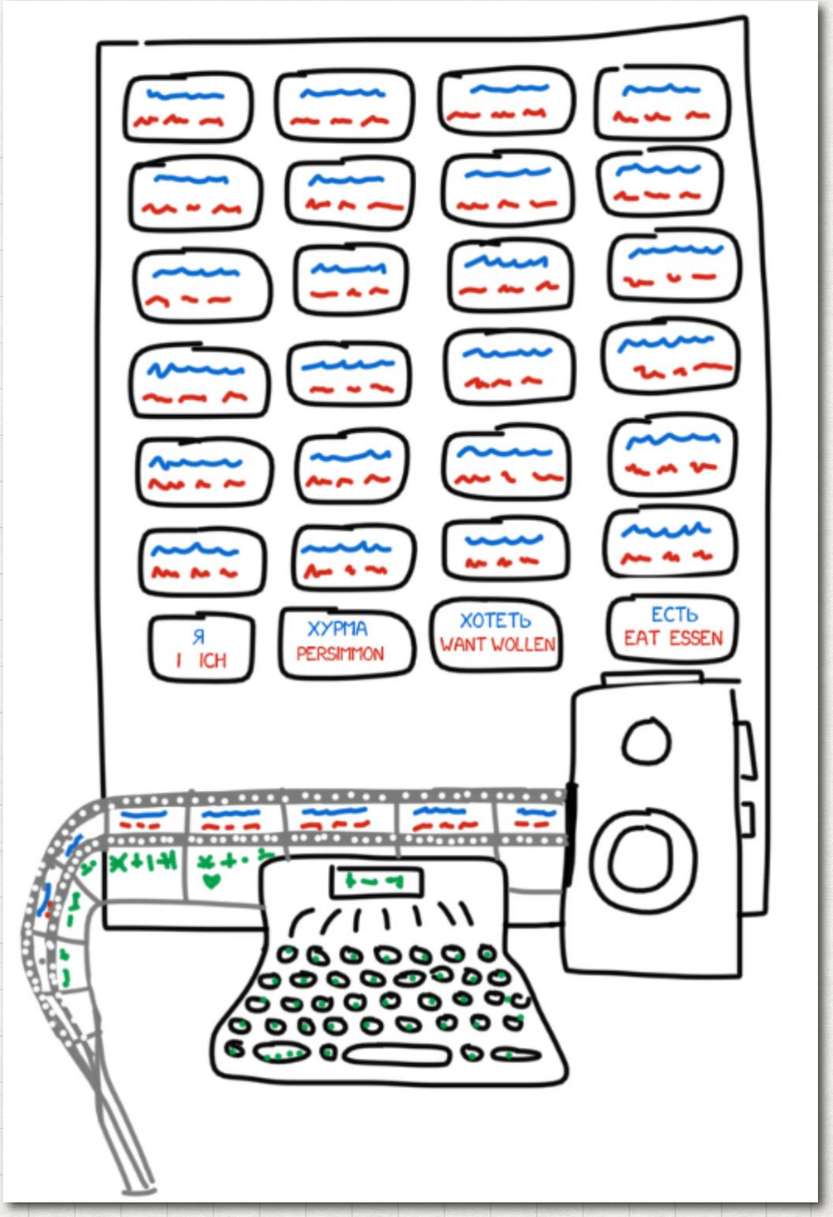
Чарльз Бэббидж

Аналитическая машин

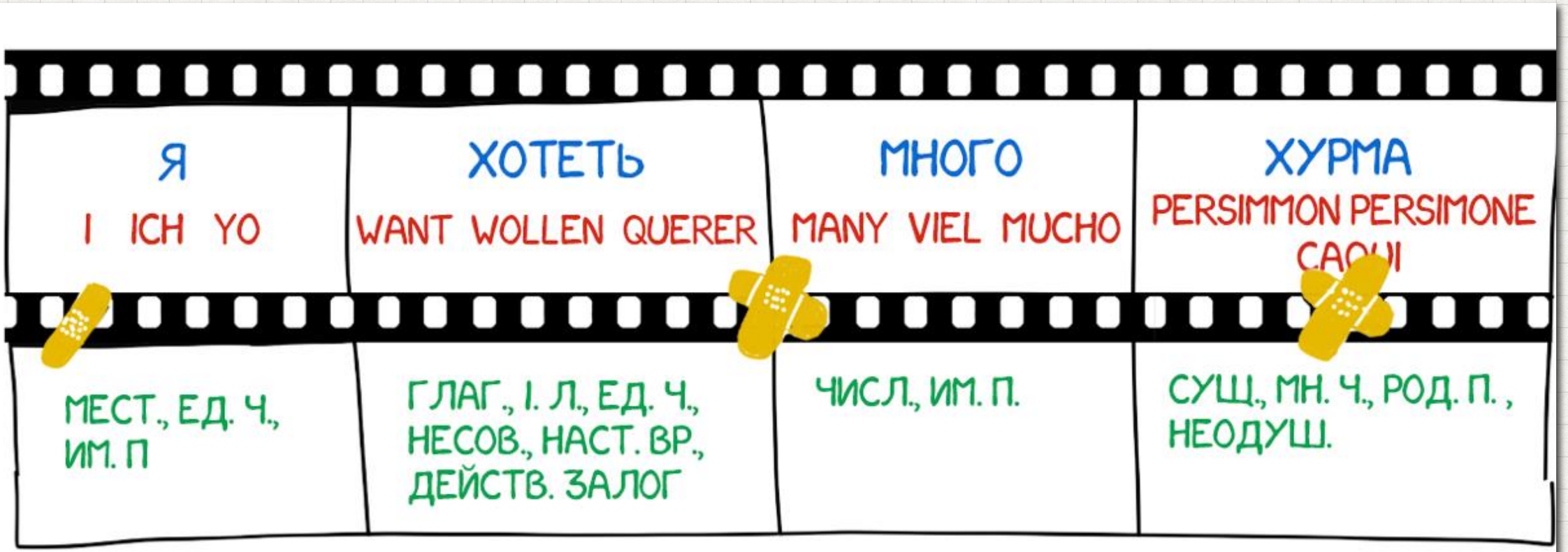
П. П. ТРОЯНСКИЙ

1933 Г.

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 с ним карточку, 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 еще Лейбниц и Декарт.


Я
I ICH YO
ХОТЕТЬ
WANT WOLLEN QUERER
МНОГО
MANY VIEL MUCHO
ХУРМА
PERSIMMON PERSIMONE
CAQUI
МЕСТ., ЕД. Ч., ИМ. П
ГЛАГ., I. Л., ЕД. Ч., НЕСОВ., НАСТ. ВР., ДЕЙСТВ. ЗАЛОГ
ЧИСЛ., ИМ. П.
СУЩ., МН. Ч., РОД. П., НЕОДУШ.

УОРРЕН УИВЕР

1947 Г.

У. Уивер (директор отделения естественных наук Рокфеллеровского фонда) написал письмо Норберту Винеру. В этом письме он предлагал использовать технику дешифрования для перевода текстов.

«У меня перед глазами текст, написанный по-русски, но я собираюсь сделать вид, что на самом деле он написан по-английски и закодирован при помощи довольно странных знаков. Все, что мне нужно, — это взломать код, чтобы извлечь информацию»

Концепция interlingva: стадия передачи информации разделена на два этапа; на первом этапе исходное предложение переводится на язык-посредник (созданный на базе упрощенного английского языка), а затем результат этого перевода представляется средствами выходного языка.

50-Е ГОДЫ:

ДЖОРДЖТАУНСКИЙ ЭКСПЕРИМЕНТ
- НЕДОЛГИЙ УСПЕХ МП

Джорджтаунский эксперимент — демонстрация возможностей машинного перевода, состоявшаяся 7 января 1954 года в Нью-Йорке, в штаб-квартире корпорации IBM.

NEW YORK, Jan. 7 (AP)— The International Business Machines Corp. put its ingenious electronic brain to work on language today and came up with a new kind of translator.

Give the brain a sentence—any old sentence such as this one in Russian:

* * *

"MYEZHDUNARODNOYE ponyimanyiye yavlyayetsya vazhnim faktorom v Ryeshyenyiyi polytyichyeskyix voprosov."

It'll be tossed back at you in English in 10 seconds.

The arrangement is mostly the doing of Dr. Leon Dostert, chairman of Georgetown University's institute of languages and linguistics, and Dr Cuthbert C. Hurd, director of IBM's applied science division.

What Dostert, Hurd and their aides have done is produce an electronic "pony"—that little book you used back in high school to help you pass your Latin course. This one's a bit larger, though.

It consists of 12 machines weighing tons each and was introduced last year by IBM as its type 701 electronic data processor. Type 701 is the rig that takes seconds to do an equation that would take you a lifetime.

* * *

JOINING IN 701'S public unveiling as a translator at IBM headquarters today was Thomas J. Watson, IBM board chairman.

"I see in this an instrumert that will be helpful in working out the problems (of world peace)," he declared. "We must do everything possible to get the people of the world to understand each other—as quickly as possible."

Dostert, who was in charge of installing the original simultaneous translation system at the United

ДЖОРДЖТАУНСКИЙ ЭКСПЕРИМЕНТ

В ходе эксперимента был продемонстрирован полностью автоматический перевод более 60 предложений с русского языка на английский.

Эксперимент был задуман и подготовлен с целью привлечения общественного и правительственного внимания.

Система МП была основана всего на 6 грамматических правилах, а словарь включал 250 записей. Система была специализированной: в качестве предметной области для перевода была выбрана органическая химия, также был добавлен ряд предложений общей тематики. Программа выполнялась на мейнфрейме IBM 701. В компьютер в торжественной обстановке на перфокартах вводились предложения вроде: «Обработка повышает качество нефти», «Командир получает сведения по телеграфу», — и машина выводила их перевод, напечатанный транслитом.

В СССР

В том же году первый эксперимент по машинному переводу был произведён в СССР, в Институте точной механики и вычислительной техники АН СССР, на компьютере БЭСМ. Руководила исследованиями Изабелла Бельская, инициировал его директор института Дмитрий Панов. Параллельно с группой Бельской над той же проблемой трудились группа ученых Отделения прикладной математики Математического института АН СССР имени В.А. Стеклова под руководством Ольги Кулагиной и Алексея Ляпунова.

А. А. Ляпунов с самого начала работ по машинному переводу говорил о переводе путем извлечения смысла переводимого текста и его представления на другом языке. Однако такая постановка проблемы перевода оказалась в то время преждевременной

КРИТИКА ДЖОРДЖТАУНСКОГО ЭКСПЕРИМЕНТА

Первое поколение систем машинного перевода базировалось на алгоритмах последовательного перевода "слово за словом", "фраза за фразой". Возможности таких систем определялись доступными размерами словарей, прямо зависящими от объема памяти компьютера. Перевод текста осуществлялся отдельными предложениями, смысловые связи между ними никак не учитывались. Такие системы называют системами прямого перевода. На смену им со временем пришли системы последующих поколений, в которых перевод от языка к языку осуществлялся на уровне синтаксических структур. В алгоритмах перевода использовался набор операций, позволяющий путем анализа переводимого предложения построить его синтаксическую структуру по правилам грамматики языка входного предложения (так же, как учат детей языку в средней школе), а затем преобразовать ее в синтаксическую структуру выходного предложения и синтезировать выходное предложение, подставляя нужные слова из словаря. Такие системы называются Т-системами (Т - от английского слова "transfer - преобразование").

«Знаете, что такое Джорджтаунский эксперимент? — спрашивает один из создателей современной системы машинного перевода «Кросслейтор» Эдуард Клышинский. — Чистое жульничество. Представьте себе словарь из 250 слов и аналитический аппарат из 6 правил. Простейшие фразы, соответствующим образом составленные, вы переведете как по маслу. Но возьмите словарь в 1000 слов. Это будет не в 4 раза. По мере увеличения словаря сложности будут нарастать экспоненциально. За каким-то порогом вы вообще не получите результата. Текста просто не будет.Машина в Джорджтауне понятия не имела, где в предложении подлежащее, а где сказуемое. Напрочь не различала ни одной формы слова и ни одной части речи. Фразы составлялись так, чтобы достаточно было простых соответствий: «мама» — «mother». Собственно, в этом и состояло невинное кибернетическое жульничество».

АМЕРИКАНСКИЙ ФИЛОСОФ ИЕГОШУА БАР-ХИЛЛЕЛ

John was looking for his toy box. Finally he found it. The box was in the pen».

«Джон искал свою коробку с игрушками. Наконец он ее нашел. Коробка была в манеже».

Автор фразы Иегошуа Бар-Хиллел заявил, что для слова «pen» («ручка», но оно же и «детский манеж») ни один электронный переводчик никогда не сможет подобрать точный аналог на другом языке. Выбор между «ручкой» и «манежем» можно сделать только имея определенную картину мира, которой у машины нет. По мнению Бар-Хиллела, этот факт закрывал тему электронного перевода навечно.

В 1967 году специально созданная комиссия Национальной Академии Наук США признала «машинный перевод нерентабельным», что существенно затормозило исследования в этой области, и рекомендовала больше сфокусироваться на разработке словарей, чем на машинном переводе, из-за чего исследователи из США практически на десятилетие выбывают из гонки.

Только в 1980-е гг. тема МП в силу удешевления машинного времени и развития технологий снова оказалась в фокусе внимания.

СССР

1956-1976 ГГ.

Один из основателей направления МП математик В. А. Успенский в своих воспоминаниях назвал эти годы "серебряным веком" структурной, прикладной и математической лингвистики в СССР.

С 1960 г. началась подготовка кадров в области автоматической переработки текстов на филологическом факультете МГУ, в Ленинградском и Новосибирском университетах, МГПИИЯ. Под математической лингвистикой понималось изучение языка как абстрактной знаковой системы с целью построения теоретической основы машинного перевода и создания конкретных алгоритмов перевода. В таком понимании математическая лингвистика составляла часть семиотики - общей теории знаковых систем.

1980-Е ГГ.

В 70-х годах разработку основ технологии машинного перевода продолжила группа специалистов в ВИНИТИ под руководством профессора Г. Г. Белоногова. В результате в 1993 г. была создана промышленная версия системы RETRANS фразеологического машинного перевода с русского языка на английский и обратно, которая применялась в министерствах обороны, путей сообщения, науки и технологий, а также во ВНТИЦ.

Первые коммерческие продукты машинного перевода, нашедшие практическое использование, появились в середине 80-х годов. Они были реализованы на персональных компьютерах и являлись системами прямого перевода, возможности которых базировались на огромных (по сравнению с первыми системами) словарях, а не на умении анализировать и синтезировать тексты.

1990-Е ГГ.

Коммерческие продукты машинного перевода предлагали отечественные фирмы:

- "Виста Текнолоджиз" и "Адвентис", образованные в 1991 г. коллективом разработчиков, выделившихся из ВИНИТИ (система МП русс-англ., электронные словари);
- ПРОМТ, образованная в 1991 г. - PROMT Internet Translation Server, электронные специализированные словари;
- "Медиа Лингва" - электронные словари МультиЛекс.

ПРОМТ

PROMT Internet Translation Server, обеспечивающий перевод "на лету" Web-страниц, запросов к поисковым системам или к базам данных, представленным в Интернете.

Для корпоративных сетей многонациональных корпораций фирма ПРОМТ предлагает аналогичный продукт PROMT Intranet Server.

Модуль перевода PROMT Internet встраивается в браузер Microsoft Internet Explorer, образуя средство для синхронного перевода Web-страниц Web View. При этом можно устанавливать для перевода различные языковые пары: английский - русский; английский - немецкий; английский - испанский; французский - английский; французский - немецкий. Бесплатный онлайновый переводчик PROMT доступен в Интернете по адресу: http://www.translate.ru.

PROMT Internet Translator Server установлен на поисковой системе Voila, принадлежащей оператору France Telecom.

70-Е И 80-Е ГГ. ЗА РУБЕЖОМ

Комиссия Европейских общин (CEC) покупает англо-французскую версию Systran, а также систему перевода с русского на английский.

CEC заказывает разработку французско-английской и итальянско-английской версий.

Проекты EUROTRA, основанного на разработках групп SUSY и GETA.

Расширение деятельности по созданию систем МП в Японии (главным образом основанных на технологии interlingva, разработанной У. Уивером в конце 40-х гг.).

В США Панамериканская организация здравоохранения (PAHO) заказывает разработку испанско-английского направления (система SPANAM).

ВВС США финансируют разработку системы МП в Лингвистическом исследовательском центре при Техасском университете в Остине.

Группа TAUM в Канаде достигает заметных успехов в разработке своей системы METEO.

КЛАССИФИКАЦИЯ МП ПО ЛАРРИ ЧАЙЛДСУ

В 1990 году Ларри Чайлдс (Larry Childs), специалист по машинному переводу, предложил следующую классификацию СМП по степени их «самостоятельности»:

- 1. FAMT (Fully-automated machine translation) - полностью автоматизированный машинный перевод;
- 2. HAMT (Human-assisted machine translation) - машинный перевод при участии человека;
- 3. MAHT (Machine-assisted human translation) - перевод, осуществляемый человеком с использованием компьютера.

ПАМЯТЬ ПЕРЕВОДОВ (TRANSLATION MEMORY)

ОСНОВНЫЕ ФУНКЦИИ ТМ

- Функция сопоставления (Alignment). Одно из преимуществ систем ТМ – это возможность использования уже переведенных материалов по данной тематике.

- Наличие фильтров импорта – экспорта. Это свойство обеспечивает совместимость систем ТМ с множеством текстовых процессоров и издательских систем и дает переводчику относительную независимость от заказчика.

- Механизм поиска нечетких или полных совпадений.

- Поддержка тематических словарей.

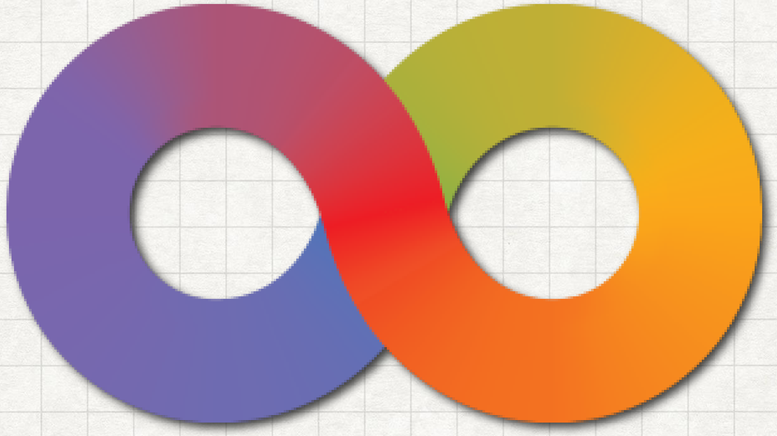
- Средства поиска фрагментов текста.

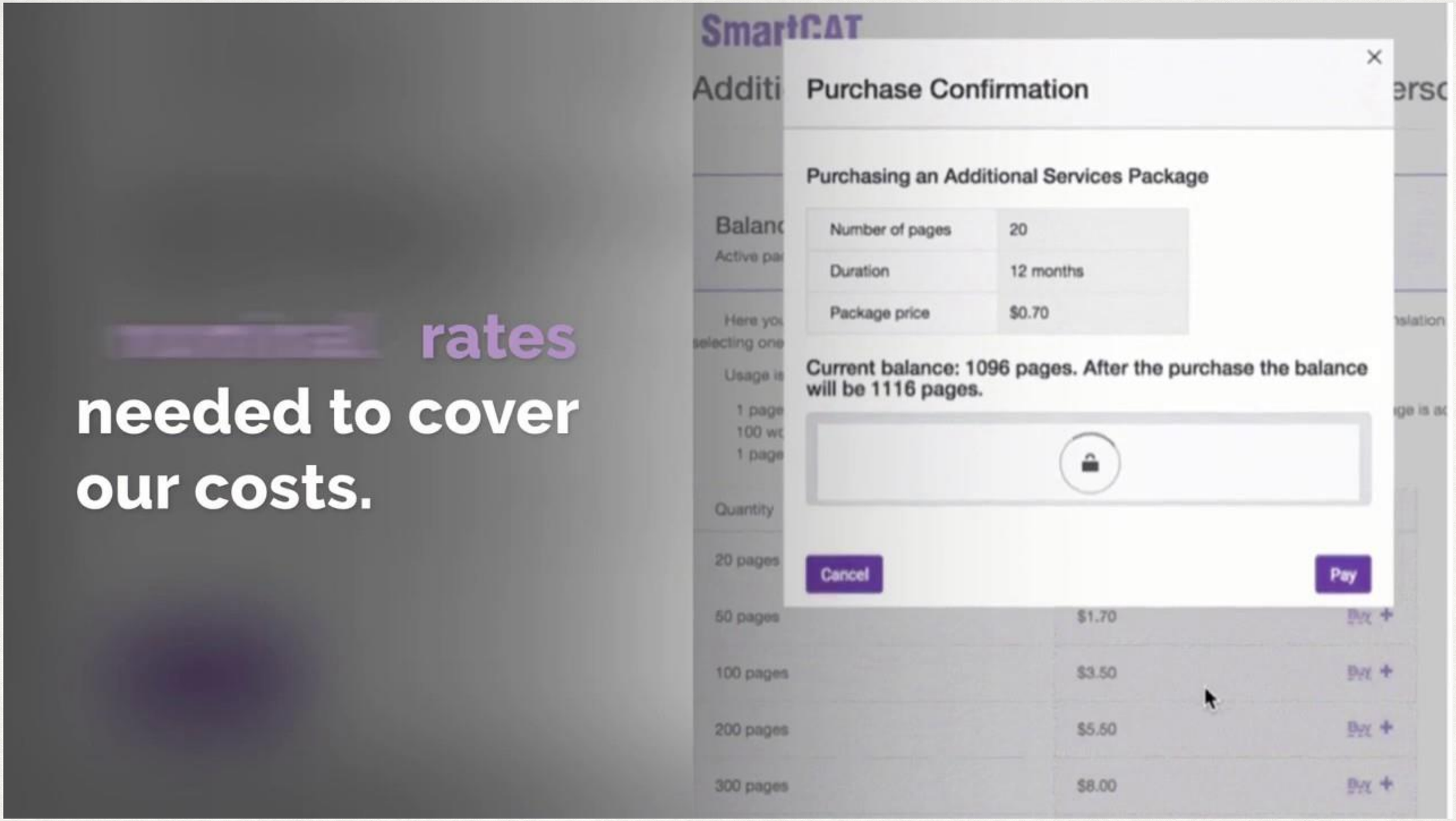
SDL TRADOS

SDL*

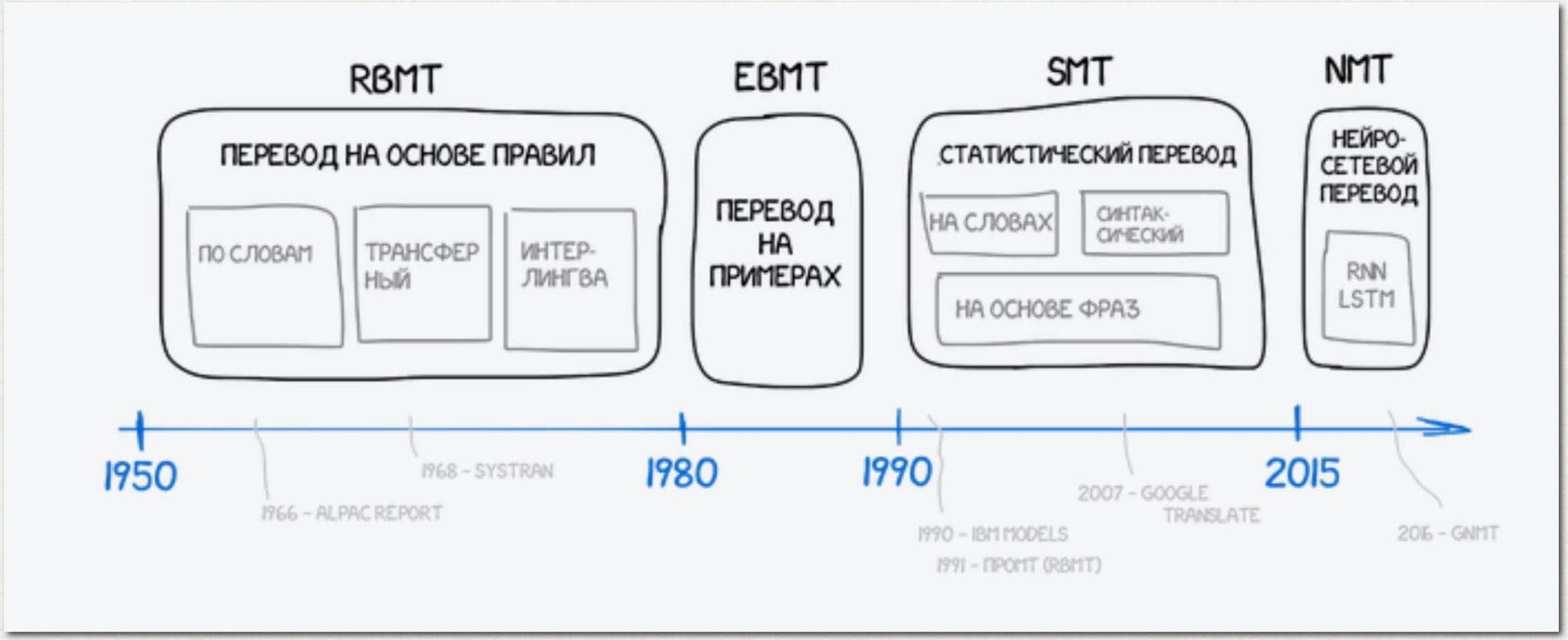
How to use neural machine translation in SDL Trados Studio

SMARTCAT




rates
needed to cover
our costs.
Purchase Confirmation
Purchasing an Additional Services Package
Number of pages 20
Duration 12 months
Package price $0.70
Current balance: 1096 pages. After the purchase the balance will be 1116 pages.
Cancel
Pay
20 pages
50 pages $1.70
100 pages $3.50
200 pages $5.50
300 pages $8.00

МЕТОДЫ МП


RBMT
ПЕРЕВОД НА ОСНОВЕ ПРАВИЛ
ПО СЛОВАМ
ТРАНСФЕР НЫЙ
ИНТЕР-ЛИНГВА
EBMT
ПЕРЕВОД НА ПРИМЕРАХ
SMT
СТАТИСТИЧЕСКИЙ ПЕРЕВОД
НА СЛОВАХ
СИНТАК-СИЧЕСКИЙ
НА ОСНОВЕ ФРАЗ
NMT
НЕЙРО-СЕТЕВОЙ ПЕРЕВОД
RNN LSTM
1950
1966 – ALPAC REPORT
1968 – SYSTRAN
1980
1990
1990 – IBM MODELS
1991 – PROMT (RBMT)
2007 – GOOGLE TRANSLATE
2015
2016 – GNMT

МАШИННЫЙ ПЕРЕВОД НА ОСНОВЕ ПРАВИЛ RULE-BASED MACHINE TRANSLATION (RBMT)

МАШИННЫЙ ПЕРЕВОД НА ОСНОВЕ ПРАВИЛ

Системы дословного перевода

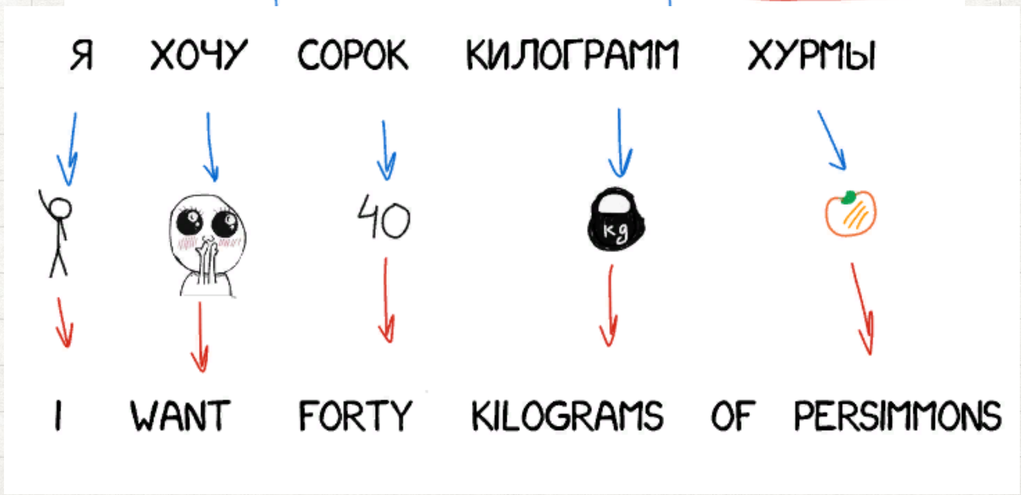
Direct Machine Translation

Трансферные системы

Transfer-based Machine Translation

Интерлингвистические системы

Interlingual Machine Translation



ПИРАМИДА ВАКУА
(VAUQUOIS)

ИНТЕРЛИНГВИСТИЧЕСКИЙ ПЕРЕВОД

ТРАНСФЕРНЫЙ ПЕРЕВОД

ДОСЛОВНЫЙ ПЕРЕВОД

ПЛЮСЫ RBMT

- синтаксическая и морфологическая точность;
- стабильность и предсказуемость результата;
- возможность настройки на предметную область.

МИНУСЫ RBMT

• трудоемкость и длительность разработки: для разработки лингвистических правил, которых в системе аналитического перевода может насчитываться десятки тысяч, необходимо привлечение лингвистов; процесс разработки этих правил может занимать от нескольких месяцев до нескольких лет;

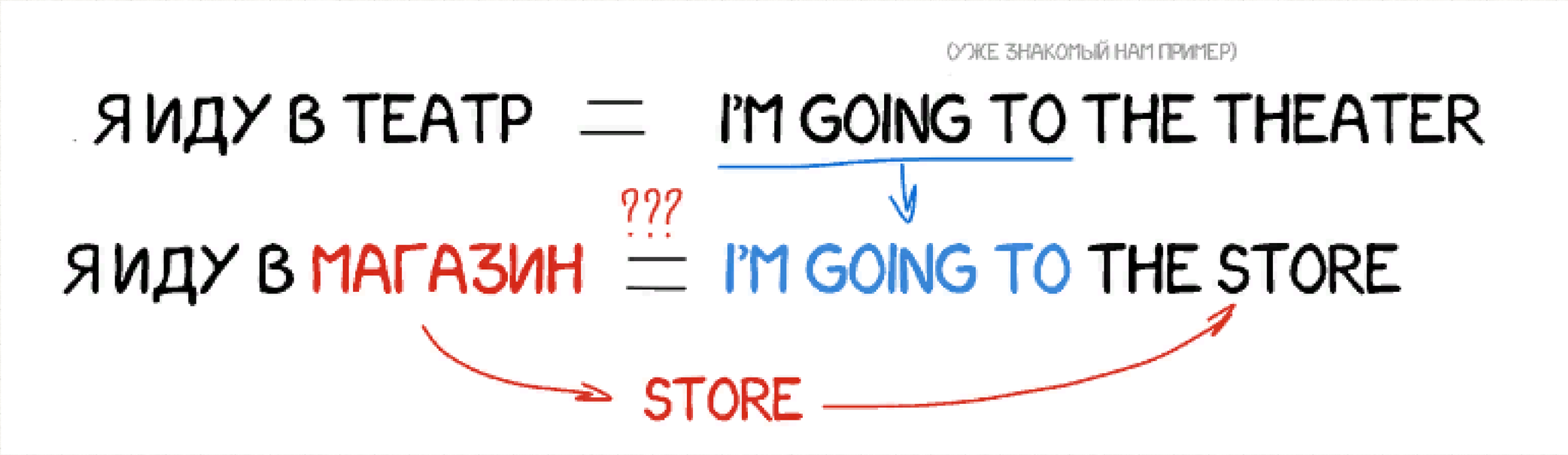
• необходимость поддерживать и актуализировать лингвистические БД;

• «машинный акцент» при переводе – ориентируясь исключительно на правила, такие системы зачастую могут игнорировать контекст, подставляя в целом верные, но не самые уместные варианты перевода отдельных лексем.

МАШИННЫЙ ПЕРЕВОД НА ПРИМЕРАХ
EXAMPLE-BASED MACHINE TRANSLATION (EBMT)

EBMT

1984 г. Университет г. Киото ученый Макото Нагато: А что если не пытаться каждый раз переводить заново, а использовать уже готовые фразы?



СТАТИСТИЧЕСКИЙ МАШИННЫЙ ПЕРЕВОД
STATISTICAL MACHINE TRANSLATION (SMT)

SMT

В 1990 году в исследовательском центре IBM впервые показали систему статистического машинного перевода.

SMT

В основе СВМТ состоит работа не с готовыми правилами, то есть не рационалистический (аналитический) подход, а эмпирический подход, то есть формирование правил на основе параллельных двуязычных корпусов текстов.

В основе технологии теорема Байеса: из предложения выделяются отдельные грамматические единицы – слова и фразы, перебираются все варианты перевода для каждого фрагмента и взвешивается вероятность каждого из них.

SMT

1. Модель перевода, или таблиц перевода – это таблица-словарь, в которой для всех известных системе слов и фраз на одном языке перечислены все возможные их переводы на другой язык и указана вероятность этих переводов.

2. Вероятностная модель языка – это представление знаний системы о языке, на который нужно перевести текст. Она используется для того, чтобы после выбора наиболее вероятных вариантов перевода отдельных слов и фраз исходя из модели перевода, выбрать из этих вариантов наиболее подходящие, исходя из контекста.

3. Декодер – составляющая переводчика, которая непосредственно переводом. Для каждого предложения исходного текста он подбирает все варианты перевода, сочетая между собой фразы из модели перевода, и сортирует их по убыванию вероятности. Затем все получившиеся варианты декодер оценивает с помощью модели языка.

ПЛЮСЫ SMT

• быстрая настройка: поскольку система обучается сама, лингвисты необходимы только для помощи в написании алгоритма анализа корпусов текстов; для дальнейшего обучения системы используются тексты, которые можно найти в свободном доступе;

• такие системы хорошо справляются с переводом сложных и редких слов, терминов и устойчивых выражений;

• легко добавлять новые направления перевода: если язык начнёт меняться, система заметит это как только к ней попадут соответствующие тексты;

• отсутствие глубокого анализа текста экономит вычислительные ресурсы.

МИНУСЫ SMT

• статистические системы гораздо хуже работают для сильно непохожих друг на друга языков без использования сложных моделей типа tree-to-tree/tree-to-string (например, при переводе с английского на японский;

• «дефицит» параллельных корпусов: качество перевода сильно зависит от количества параллельных корпусов, для корректных переводов статистической системе необходимо как минимум 500 тысяч, в идеале от нескольких миллионов параллельных текстов;

• нестабильность перевода: несмотря на способность переводить устойчивые выражения, эти самые выражения и термины могут переводиться по-разному исходя из контекста;

• зачастую результат перевода похож на «собранный паззл»: хотя общий смысл предложения понятен, но части предложений существуют отдельно друг от друга.

ГИБРИДНЫЙ МАШИННЫЙ ПЕРЕВОД
HYBRID MACHINE TRANSLATION

ГИБРИДНЫЕ СИСТЕМЫ

Гибридная технология перевода предполагает использование статистических методов для построения словарных баз автоматическим путём на основе параллельных корпусов, формирования нескольких возможных переводов как на лексическом уровне, так и на уровне синтаксической структуры предложения выходного языка, применения постредактирования в автоматическом режиме и выбор лучшего (наиболее вероятного) перевода из возможных на основе языковой модели, построенной по определенному корпусу выходного языка.

ГИБРИДНЫЕ СИСТЕМЫ

Например, гибридные системы RBMT и SMT

- Rule-based MT с пост-обработкой статистического подхода.
- Statistical MT с предварительной обработкой по Rule-based подходу.
- Полная интеграция RBMT и SMT.

ГИБРИДНЫЕ СИСТЕМЫ

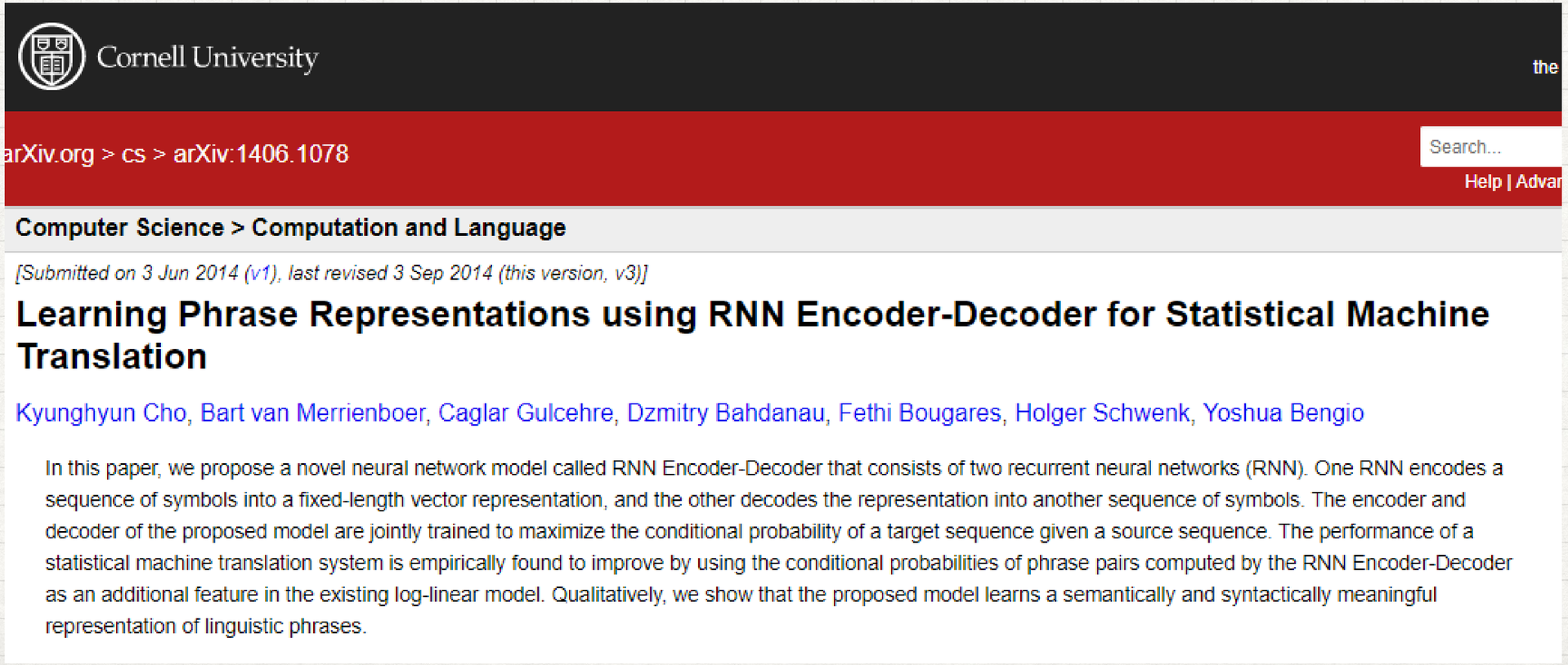
Объединение RBMT и статистического машинного перевода дает:
• Лингвистический анализ входного предложения;
• Порождение вариантов перевода;
• Использование статистических технологий;
• Оценка и выбор лучшего варианта перевода с использованием языковой модели.

Этапы гибридной технологии SMT и RBMT:
• Обучение RBMT на основе параллельного корпуса с использованием статистических технологий;
• Эксплуатация на основе натренированной системы.

НЕЙРОННЫЙ МАШИННЫЙ ПЕРЕВОД
NEURAL MACHINE TRANSLATION

NMT

В 2014 году выходит статья с кратким описанием идеи применения нейросетей глубокого обучения к машинному переводу.



Cornell University

the

arXiv.org > cs > arXiv:1406.1078

Search...

Help | Advar

Computer Science > Computation and Language

[Submitted on 3 Jun 2014 (v1), last revised 3 Sep 2014 (this version, v3)]

Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, Yoshua Bengio

In this paper, we propose a novel neural network model called RNN Encoder-Decoder that consists of two recurrent neural networks (RNN). One RNN encodes a sequence of symbols into a fixed-length vector representation, and the other decodes the representation into another sequence of symbols. The encoder and decoder of the proposed model are jointly trained to maximize the conditional probability of a target sequence given a source sequence. The performance of a statistical machine translation system is empirically found to improve by using the conditional probabilities of phrase pairs computed by the RNN Encoder-Decoder as an additional feature in the existing log-linear model. Qualitatively, we show that the proposed model learns a semantically and syntactically meaningful representation of linguistic phrases.

NMT

В ноябре 2016 г., в блоге Google появляется анонс НМП.

Google AI Blog

The latest news from Google AI

A Neural Network for Machine Translation, at Production Scale

Tuesday, September 27, 2016

Posted by Quoc V. Le & Mike Schuster, Research Scientists, Google Brain Team

Ten years ago, we announced the launch of Google Translate, together with the use of Phrase-Based Machine Translation as the key algorithm behind this service. Since then, rapid advances in machine intelligence have improved our speech recognition and image recognition capabilities, but improving machine translation remains a challenging goal.

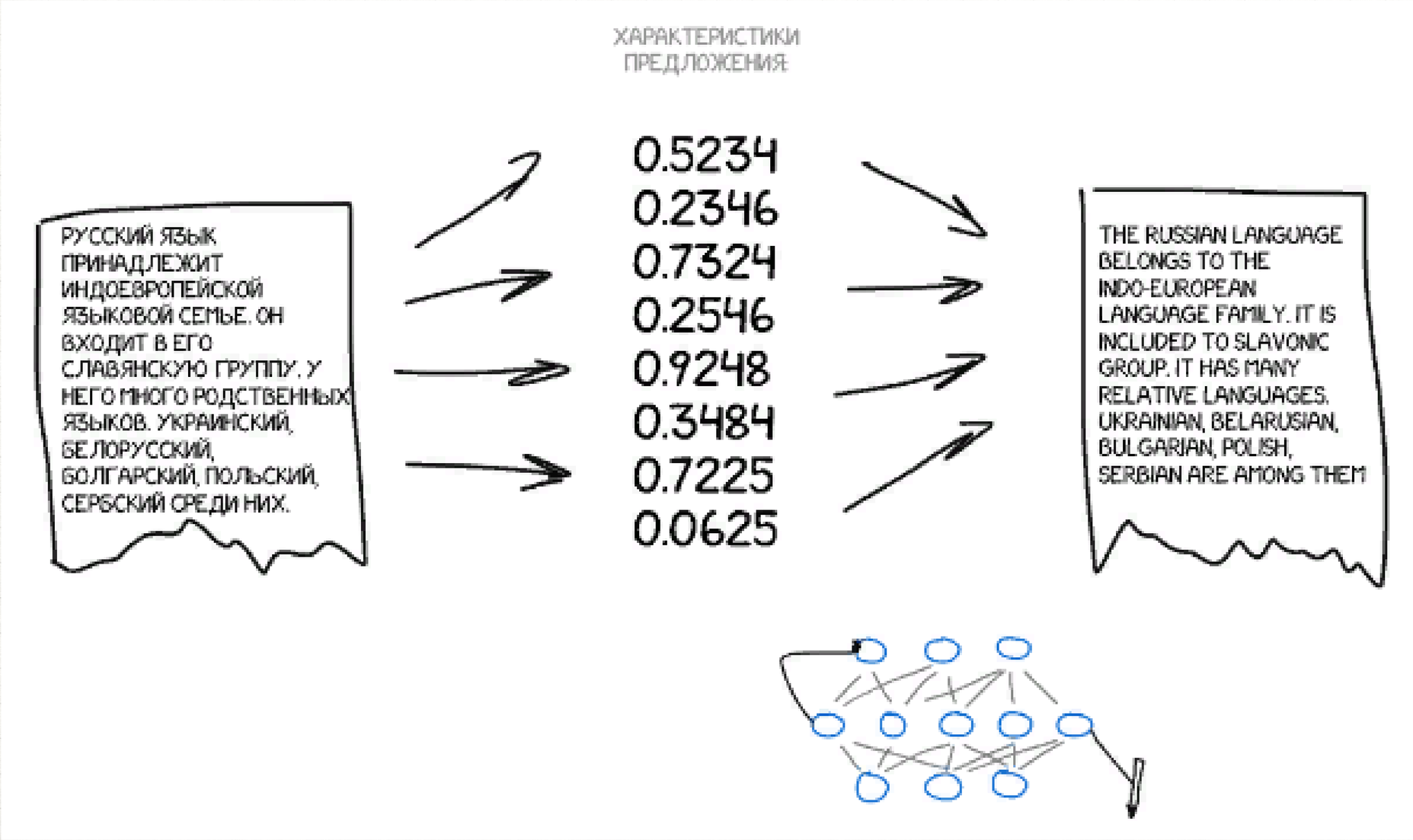
Today we announce the Google Neural Machine Translation system (GNMT), which utilizes state-of-the-art training techniques to achieve the largest improvements to date for machine translation quality. Our full research results are described in a new technical report we are releasing today: "*Google's Neural Machine Translation System: Bridging the Gap between Human and Machine*

NMT

В основе нейронного переводчика - механизм двунаправленных рекуррентных нейронных сетей (RNN), построенный на матричных вычислениях, который позволяет строить существенно более сложные вероятностные модели, чем статистические машинные переводчики.

Хотя нейронный перевод также использует для обучения параллельные корпуса, в процессе обучения он оперирует не отдельными фразами, но целыми предложениями. Одна из главных проблем состоит в том, что нейронной сети требуется гораздо больше корпусов для обучения, чем статистической системе: по меньшей мере порядка 100 миллионов токенов для адекватной работы, для переводов же надлежащего качества – не менее порядка 500 миллионов. Также для обучения подобной системы требуется куда больше вычислительных мощностей.

NMT


ХАРАКТЕРИСТИКИ ПРЕДЛОЖЕНИЯ:
РУССКИЙ ЯЗЫК ПРИНАДЛЕЖИТ ИНДОЕВРОПЕЙСКОЙ ЯЗЫКОВОЙ СЕМЬЕ. ОН ВХОДИТ В ЕГО СЛАВЯНСКУЮ ГРУППУ. У НЕГО МНОГО РОДСТВЕННЫХ ЯЗЫКОВ. УКРАИНСКИЙ, БЕЛОРУССКИЙ, БОЛГАРСКИЙ, ПОЛЬСКИЙ, СЕРБСКИЙ СРЕДИ НИХ.
0.5234
0.2346
0.7324
0.2546
0.9248
0.3484
0.7225
0.0625
THE RUSSIAN LANGUAGE BELONGS TO THE INDO-EUROPEAN LANGUAGE FAMILY. IT IS INCLUDED TO SLAVONIC GROUP. IT HAS MANY RELATIVE LANGUAGES. UKRAINIAN, BELARUSIAN, BULGARIAN, POLISH, SERBIAN ARE AMONG THEM

ГЛУБОКОЕ ОБУЧЕНИЕ

Глубокое обучение (глубинное обучение; англ. Deep learning) — совокупность методов машинного обучения (с учителем, с частичным привлечением учителя, без учителя, с подкреплением), основанных на обучении представлениям (англ. feature/representation learning), а не специализированным алгоритмам под конкретные задачи.

Примеры использования:
в компьютерном зрении, машинном переводе, распознавании речи, причём качество решения во многих случаях теперь сопоставимо, а в некоторых случаях превосходит эффективность человека.

ГЛУБОКОЕ ОБУЧЕНИЕ

Глубинная нейронная сеть (ГНС, англ. DNN — Deep neural network) — это искусственная нейронная сеть (ИНС) с несколькими слоями между входным и выходным слоями.

ГНС находит корректный метод математических преобразований, чтобы превратить исходящие данные в выходящие, независимо от линейной или нелинейной корреляции. Сеть продвигается по слоям, рассчитывая вероятность каждого выхода.

ОБЩИЕ ВЫВОДЫ

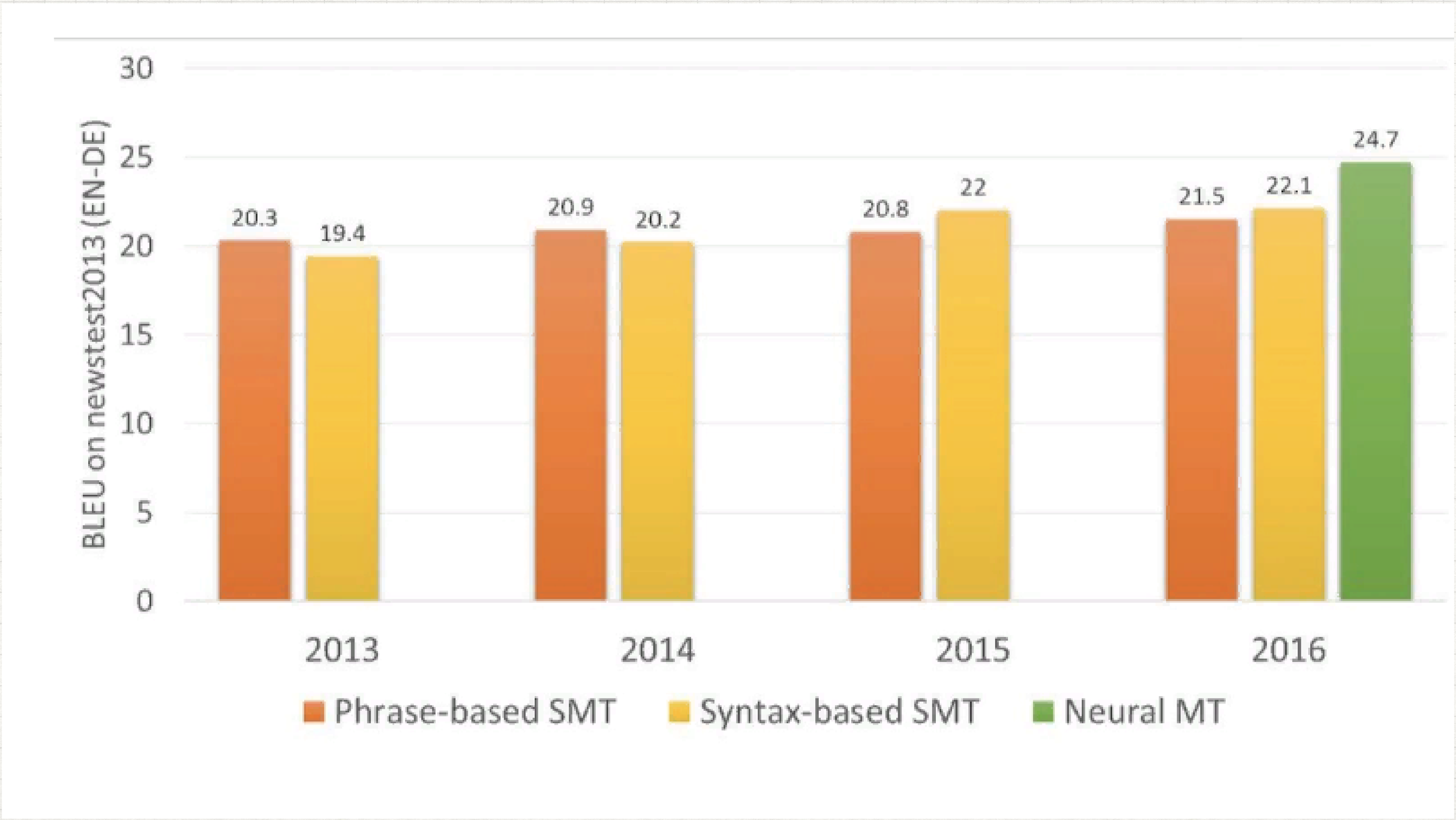
- В ОБЩЕМ СЛУЧАЕ НЕЙРОННЫЙ АВТОМАТИЧЕСКИЙ ПЕРЕВОД ДАЕТ РЕЗУЛЬТАТ БОЛЕЕ ВЫСОКОГО КАЧЕСТВА, ЧЕМ «ЧИСТО» СТАТИСТИЧЕСКИЙ ПОДХОД;

- АВТОМАТИЧЕСКИЙ ПЕРЕВОД ЧЕРЕЗ НЕЙРОННУЮ СЕТЬ ЛУЧШЕ ПОДХОДИТ ДЛЯ РЕШЕНИЯ ЗАДАЧИ «УНИВЕРСАЛЬНОГО ПЕРЕВОДА»;

- НИ ОДИН ИЗ ПОДХОДОВ К МАШИННОМУ ПЕРЕВОДУ САМ ПО СЕБЕ НЕ ЯВЛЯЕТСЯ ИДЕАЛЬНЫМ УНИВЕРСАЛЬНЫМ ИНСТРУМЕНТОМ ДЛЯ РЕШЕНИЯ ЛЮБОЙ ЗАДАЧИ ПЕРЕВОДА;

- ДЛЯ РЕШЕНИЯ ЗАДАЧ ПО ПЕРЕВОДУ В БИЗНЕСЕ ТОЛЬКО СПЕЦИАЛИЗИРОВАННЫЕ РЕШЕНИЯ МОГУТ ГАРАНТИРОВАТЬ СООТВЕТСТВИЕ ВСЕМ ТРЕБОВАНИЯМ.

Оценка качества МП



СТРАТЕГИИ ВЗАИМОДЕЙСТВИЯ ЧЕЛОВЕК–МАШИНА ПРИ ПЕРЕВОДЕ

– компьютерный перевод с предредактированием – это преобразование текста перед его вводом в систему компьютерного перевода для воссоздания на языке оригинала конструкций языка перевода;
– компьютерный перевод с постредактированием – это преобразование «грубого» текста, выполненного системой машинного перевода, с целью привести его к нормам языка перевода;
– компьютерный перевод с интерредактированием – это взаимодействие человека и машины непосредственно во время перевода.

РЕДАКТИРОВАНИЕ МП

- простое;
- полное.

ПРОСТОЕ РЕДАКТИРОВАНИЕ

заключается в проведении как можно меньшего числа операций над текстом, с целью сделать его понятным, фактически точным и грамматически правильным и включает в себя:

– исправление наиболее очевидных опечаток и грамматических ошибок;

– изменение сложных предложений частично или полностью;

– фиксирование ошибок компьютерного переводчика;

– удаление ненужных или альтернативных вариантов перевода;

– создание глоссария, но без углубленной проработки терминов.

Этот вид редактирования применяется в тех случаях, когда необходимо передать только смысл

ПОЛНОЕ ПОСТРЕДАКТИРОВАНИЕ

Полное постредактирование – это более долгий процесс, итогом которого является текст, который читается, как будто он был написан на языке перевода. Этот процесс включает в себя:

– проверку соответствия терминологии;
– сбор информации, связанной с текстом перевода;
– синтаксические изменения в соответствии с правилами языка перевода;
– работа над стилистикой текста;
– перевод и адаптация культурологических явлений (фразеологизмы, идиомы и др.);
– попытка добиться полного соответствия с оригинальным текстом;
– выполнение форматирования, в соответствии с оригиналом текста;
– исправление всех грамматических, пунктуационных и орфографических ошибок.

ОЦЕНКА КАЧЕСТВА МП

- на основе лингвостилистического анализа;
- на основе автоматической обработки текстов.

ОЦЕНКА КАЧЕСТВА МП НА ОСНОВЕ ЛИНГВОСТИЛИСТИЧЕСКОГО АНАЛИЗА

Макото Нагао предложил шкалу оценки машинного перевода на основе лингвостилистического анализа, от пяти до одного балла.

- 1 балл – смысл предложения понятен и не возникает никаких вопросов, грамматика, словоупотребление и стиль соответствуют общей структуре текста и не требуют постредактирования
- 2 балла – смысл предложения понятен, но возникают большие проблемы с грамматикой, словоупотреблением и стилем.
- 3 балла – общий смысл предложения понятен, но смысл некоторых его частей вызывает сомнение из-за неправильного грамматического строя.
- 4 балла – присутствуют ошибки словоупотребления и стилистики, требуется обращение к оригиналу.
- 5 баллов – в предложении имеется большое количество грамматических, словоупотребительных и стилистических ошибок, смысл предложения с трудом можно понять после внимательного изучения

АВТОМАТИЧЕСКИЕ СИСТЕМЫ ОЦЕНКИ МАШИННОГО ПЕРЕВОДА

основываются на методе N-грамм, который был введен для этой цели также Макото Нагао и Шинсуке Мори. Этот метод основан на использовании вероятности появления цепочки букв N-го порядка (N-грамм) в анализируемых текстах. N-грамма – это последовательность из n-слов или знаков. По своему составу N-граммы могут делиться на униграммы (одно слово или один знак), биграммы (два), триграммы (четыре) и так далее.

Основными метриками, основанными на этом методе, являются BLEU и METEOR (IBM).

АВТОМАТИЧЕСКИЕ СИСТЕМЫ ОЦЕНКИ МАШИННОГО ПЕРЕВОДА

Алгоритм оценивания следующий:
в программу загружаются два перевода одного и того же текста.
Один, выполненный с помощью системы компьютерного перевода, второй, прошедший постредактирование.
Оба этих текста делятся на N-граммы, после чего сравниваются друг с другом.
Далее, используя методы математической статистики, высчитывается оценка компьютерного перевода. Оценка ставится от 0 до 1, где 0 – плохой результат, а 1 – отличный.