Машинный перевод

Ефремова Наталья Эрнестовна
Грацианова Татьяна Юрьевна
Большакова Елена Игоревна

Содержание

- Сложность задачи
- История. Поколения систем перевода
- Технологии машинного перевода (МП)
  - Перевод на основе правил
  - Статистический перевод
- Методы оценки качества МП
- Некоторые системы МП
- Сравнение систем и переводов
- Заключение: реальность и перспективы

Sonnet 90
by William Shakespeare

Then hate me when thou wilt, if ever, now	Тогда ненавидьте меня, когда Вы будете слабеть, если когда-либо, теперь
Now while the world is bent my deeds to cross,	Теперь, в то время как мир согнут мои дела на крест,
Join with the spite of Fortune, make me bow,	Соединение со злостью Fortune, заставьте меня поклониться,
And do not drop in for an after-loss.	И не бросать! в за после-того,-как-потерю.
Ah do not, when my heart has scaped this sorrow,	Ах не делайте, когда мое сердце иссушит это горе,
Come in the rearward of a conquered woe;	Войдите назад завоеванного горя;
Give not a windy night a rainy morrow,	Дайте не ветреной ночи дождливый следующий день,
To linger out a purposed overthrow.	Задерживаться ставившее целью свержение.
If thou wilt leave me, do not leave me last,	Если Вы слабеете, оставляют меня, не оставляйте меня в последний раз,
When other petty griefs have done their spite,	Когда другое мелкое горе сделало свою злость,
But in the onset come; so shall I taste	Но в начале прибывает; так буду я являться на вкус
At first the very worst of Fortune's might;	Сначала очень худший из силы Fortune;
And other strains of woe, which now seem woe,	И другие напряжения горя, которые теперь кажутся горем,
Compared with loss of thee, will not seem so.	По сравнению с потерей тебя, не будет казаться так.

Литературный перевод «Сонет 90»

Уж если ты разлюбишь - так теперь,
Теперь, когда весь мир со мной в раздоре.
Будь самой горькой из моих потерь,
Но только не последней каплей горя!

С.Маршак

Коль хочешь быть врагом мне - будь им ныне,
Когда гоним я роком и людьми.
Примкни скорей к враждебной мне судьбине,
Но медленной отравой не томи!

М.Чайковский

Что ж, ненавидь, коль хочешь! Но сейчас,
Сейчас, когда грозит мне злобой небо.
Согни меня, с судьбой объединясь,
Но лишь бы твой удар последним не был.

А.Финкель

Необходимость Машинного Перевода (МП)

- свыше 7 млрд жителей Земли используют около 7 тыс. языков, и все большее их количество включается в мировые информационные потоки
- Европейский союз в настоящее время объединяет 27 государств, в которых используется 23 официальных языка
- Подсчитано (80-е годы), что если синтез нового химического соединения обойдется менее чем в 100 тыс. долларов, выгоднее произвести этот синтез, чем искать описание аналогичной работы на других языках

История Машинного Перевода (МП)

- 40-е – идея применения компьютеров для МП
- 1952 г. – Джорджтаунский эксперимент
- 60-е – прекращение финансирования
- 70-80-е – «ренессанс»
- 90-е – развитие на базе новых технологий

Идея применения компьютеров для МП

I have a text in front of me which is written in Russian but I am going to pretend that it is really written in English and that it has been coded in some strange symbols. All I need to do is strip off the code in order to retrieve the information contained in the text

Warren Weaver март 1947 г.

Джорджтаунский эксперимент

1954 г.
-С русского языка на английский
- 49 предложений
- словарь 250 слов
- 6 грамматических правил

Идея о принципиальной невозможности правильного автоматического перевода – 1959 г.

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

Джон искал свою игрушечную коробку. Наконец он её нашёл. Коробка была в манеже. Джон был очень счастлив.

1966 г. – заключение ALPAC о нерентабельности МП

Первое поколение систем МП

- 50-60 гг. – двуязычные системы,
  простейшие лингвистические модели: последовательный перевод,
  пословный и *пословно-пооборотный* перевод –
  приемлемое качество только для родственных языков, например:
  испанский-португальский

- 60-70 гг. – *пофразный* перевод, применение синтаксического анализа,
  стратегия АНАЛИЗ $\Rightarrow$ ТРАНСФЕР $\Rightarrow$ СИНТЕЗ
  – модульность (грамматика и словарь)
  – пред- и пост-редактирование человеком
  – появление промышленных систем: *SYSTRAN* – США, 1970 гг. :
  перевод научно-технических текстов

- Первые работы в России,1955 г.:
  – математические тексты с английского на русский, словарь 2300
  слов, ИТМ и ВТ АН СССР, Трифонов Н.П., Королев Л.Н.

  – работы в ИПМ им. Келдыша (О.С. Кулагина)

Ренессанс

70-е - 80-е годы

Смещение акцентов:

- развитие коммерческих, «реалистичных» систем
- автоматизированный перевод
- многоязычные системы

Новые идеи:

- идея внутреннего универсального семантического *языка-посредника*
- японский проект «ЭВМ 5-го поколения», технология *interlingva*
- технология *TM (translation memory)*

Виды МП с точки зрения роли человека в процессе

- МАНТ *(Machine-assisted human translation)* — перевод, осуществляемый человеком с использованием компьютера;
- НАМТ *(Human-assisted machine translation)* — машинный перевод при участии человека;
- FAMT *(Fully-automated machine translation)* — полностью автоматизированный машинный перевод.

Технология *TM*

Программа автоматизации перевода, управляющая оболочка, работающая с базой, «копилкой переводов».

База : билингвы – пары соответствующих друг другу фраз на языке оригинала и на языке перевода.

Совпадения части переводимого текста с фрагментами из базы подсвечивается.

База пополняется в процессе перевода текстов

Проблемы большой базы: долгий поиск, противоречивость

Достоинства: одинаковый перевод терминов; рассматривая все билингвы из базы с данным термином можно построить качественный перевод

Современные методы ТМ основаны на нейронных сетях, распознают словоизменение и нарушение порядка слов

Системы, основанные на технологии ТМ

- Translation Manager (IBM)
- SDLX (SDL, Великобритания)
- Déjà Vu (Atril Language Engineering, Испания)
- Translator’s Workbench, Trados (Trados GmbH, Германия)

Россия. Словари, формируемые сообществом, содержат словосочетания, фразы

http://www.multitran.ru, http://bab.la

Пример перевода:

first in first out algorithm

- алгоритм последовательного обслуживания

СРЕДНЕЕ ПОКОЛЕНИЕ

80-е – 90-е

- Россия (организации и отечественные системы):
  - ВЦП: англо/немецко/французско-русский перевод – системы *АМПАР, НЕРПА, ФРАП*
  - ИнформЭлектро / ИППИ: система *ЭТАП* французско/английско-русский перевод научно-технических текстов, модель ЕЯ «Смысл⇔Текст» (одна из наиболее полных лингвистических моделей МП)
- Канада: с 1976 г. *TAUM METEO* – полноценная система перевода метеосводок с английского на французский, использована модель«Смысл⇔Текст»

Теория „Смысл ⇔ Текст“

Создана Мельчуком А.А., Жолковским А.К., Апресяном Ю.Д. в середине 60-х годов

Многоуровневая модель преобразования

Смысл → Текст и Текст →Смысл

Естественный язык понимается как система правил, обеспечивающая этот переход.

Уровни языка:

- фонологический (текст)
- поверхностно-морфологический
- глубинно-морфологический
- поверхностно-синтаксический
- глубинно-синтаксический
- семантический (уровень смысла)

Третье поколение – современные системы МП

Прогресс в конце 90-х:

- ✓ Бурный рост объемов информации, возможность ее хранения и обработки
- ✓ Рост скорости обработки информации
- ✓ Развитие интернета
- ✓ Появление персональных гаджетов различного назначения и возможностей

Это обеспечило новые возможности для развития МП (хранение словарей, накопление текстов, обработка больших корпусов текстов и повысило потребности пользователей в системах МП. Совершенствуются старые технологии, вырабатываются новые.

Технологии МП

- Rule-based Machine Translation (RBMT), Машинный перевод, основанный **на правилах**

- Example-Based MT (EBMT), Машинный перевод, основанный **на примерах**

- Statistical Machine Translation (SMT), **Статистический** машинный перевод

- Hybrid Machine Translation (HMT), **Гибридный** машинный перевод

- Neural Machine Translation **Нейронный** машинный перевод (NMT)

Перевод на основе правил: виды

- Пословный перевод
  слова входного текста → слова выходного текста

- Трансферные системы: морфологический, синтаксический и семантический анализ на языке входа; преобразование в структуру выходного языка (TRANSFER); синтез на выходном языке

исходное предложение →
препарированное исходное предложение →
предложение на другом языке

- Интерлингвистические системы: анализ и синтез через язык-посредник

исходный текст →
описание его смысла на языке-посреднике →
текст перевода

Трансферный подход

1. Токенизация
2. Анализ токенов (часть речи, грамматическое значение)
3. Сопоставление «слово из предложения – словарная статья»
4. Анализ структуры предложения (связи между словами, объединение слов в группы)
5. Выбор варианта перевода слова
6. Построение предложения на конечном языке

Пример (трансфер)

PROMT (ранние версии)

PROgrammer’s Machine Translation

- Система словарей:
  - Общелексический словарь
  - Специализированные (отраслевые) словари
  - Словари основ: базовая форма, псевдооснова, грамматические и семантические признаки, переводы с пометами
  - Таблицы флексий (все типы изменения слов)
  - Вспомогательные таблицы: базы префиксов и постфиксов, имен и географических названий

- Модуль перевода
  - Грамматические правила (десятки тысяч)
  - Алгоритмы перевода
  - Управление стилем перевода
  - Контроль качества перевода (сравнение вариантов)

Пример (трансфер и интерлингва)

Система Этап http://proling.iitp.ru/ru/etap3
(Институт Передачи Информации РАН)

ЭТАП – 1, ЭТАП-2 - трансфер

Несколько ЕЯ пар:

русско-немецкий/французский/испанский/корейский
англо-русский/арабский

Переход с одного языка на другой через синтаксическую структуру, деревья зависимостей

Модель «Смысл⇔Текст»

В последние годы (ЭТАП-3) использует *UNL* – универсальный сетевой язык

Система Этап-3

- Англо-русский и русско-английский словари (100 тыс. лексических единиц каждый)
- Синтаксический анализ – построение дерева зависимостей слов в предложении (имена и наборы характеристик)
- Массив правил анализа, синтеза и перевода текстов
- Лексические функции для описания нестандартной синтактики

Magn (disease) = grave *Magn (болезнь) = тяжелая*
Magn (fog) = heavy *Magn (туман) = густой*

- Конвертор/деконвертор семантического языка UNL

Возможности системы ЭТАП

- Выбор тематики текста, возможность самонастройки
- Множественный перевод, синонимическое перефразирование:
  - *They made a general remark that …*
  - *They remarked in a general way that …*
  - *They forced a general to remark that …*
- Разрешение неоднозначности:
  - лексической: *to draw a distinction*
    (больше 50 значений для *to draw* ,
    больше 10 значений глагола *проводить*)
  - синтаксической: *support of the parliament*
    ? *support by the parliament*
    Vs. support (given) to the parliament

Перевод на основе правил: особенности

Достоинства:

- Синтаксическая и морфологическая точность
- Стабильность и предсказуемость результата
- Возможность настройки на предметную область

Недостатки:

- Трудоемкость и длительность разработки
- Необходимость поддерживать и актуализировать лингвистические базы данных
-«Машинный акцент» – неестественность перевода

Перевод, основанный на примерах

Требуется БЗ - языковой корпус из пар предложений.

Перевод по аналогии

Английский	**Японский (латиница)**
*How much is that **red umbrella**?*	*Ano **akai kasa** wa ikura desu ka.*
*How much is that **small camera**?*	*Ano **chiisai kamera** wa ikura desu ka*

Достоинства - высокое качество перевода, хорошо справляется с контекстными задачами (перевод идиом), для разработки не нужны квалифицированные лингвисты.

Недостатки - нужны большие параллельные корпуса текста, размеченные определенным образом, от них сильно зависит перевод, продолжительное время обучения, требовательность к ресурсам

Статистический перевод: основы

БАЗА

Параллельные текстовые корпусы:

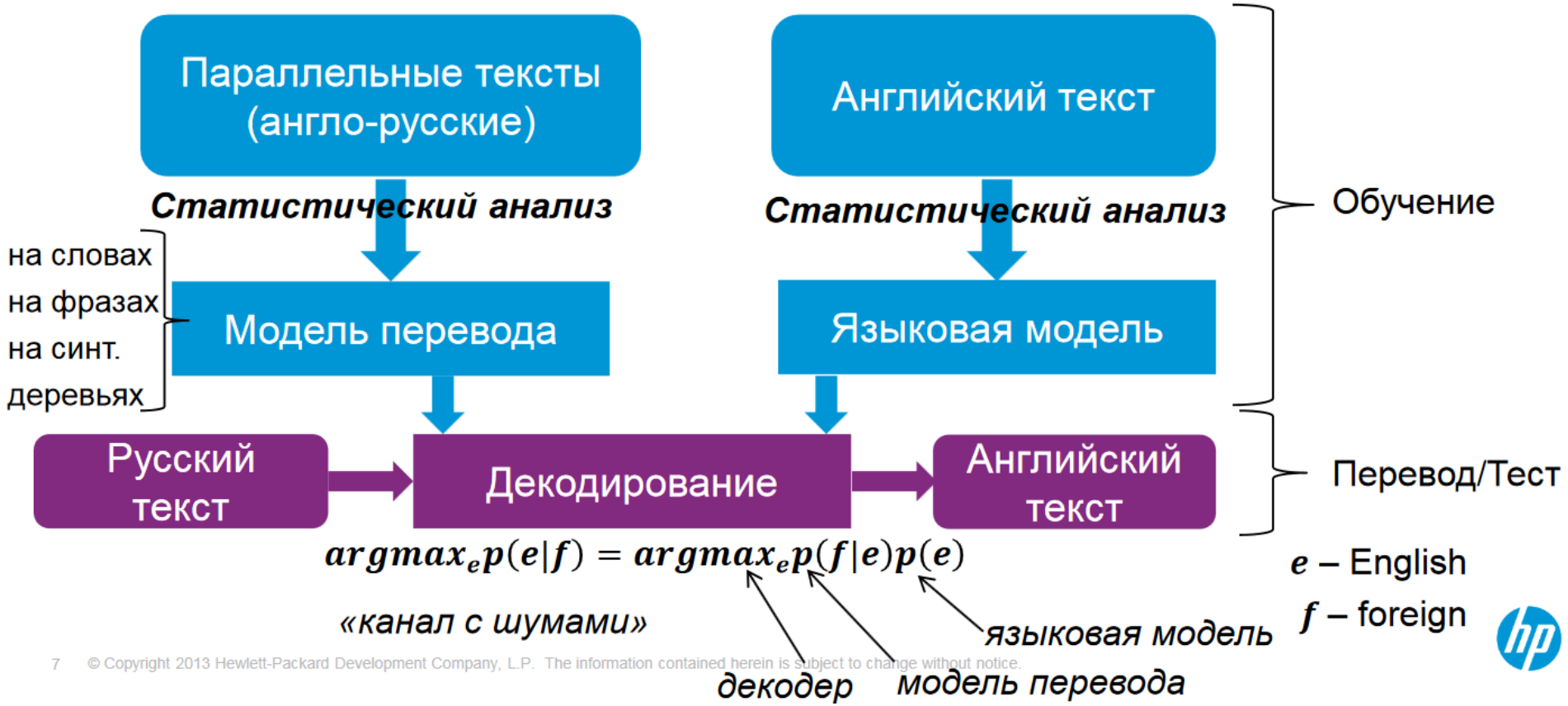
парламентские отчеты, документы Евросоюза, переводы роликов и фильмов

ТЕОРЕТИЧЕСКАЯ ОСНОВА

Статистические языковые модели
(при переводе учитываются наиболее частотные словоупотребления документа).

Языковая модель – набор n-грамм многоязычного корпуса с их вероятностными характеристиками, на базе которого ведется поиск наиболее вероятного перевода.

Статистический машинный перевод: схема



Формула перевода

Идея: перевод = дешифровка

Задача: найти в конечном языке такое предложение Y, которое с наибольшей вероятностью является переводом предложения X начального языка. P(y|x)

X = This cat is nice

Y1= Этот кошка хорош	P(Y1|X)
Y2 = Эта этот кот	P(Y2|X)
Y3 = Эта кошка хороша	P(Y3|X)
Y4 = Эта кошка есть красивый	P(Y4|X)
Y5 = Этот кот красивый	P(Y5|X)
Y6= Вася – дурак	P(Y6|X)

$$\operatorname{argmax}_y \; P(Y|X)$$

Преобразование формулы с помощью теоремы Байеса

$$argmax_y P(Y|X) = argmax_y \frac{P(Y) \cdot P(X|Y)}{P(X)}$$

$$argmax_y P(Y|X) = argmax_y P(Y)P(X|Y)$$

P(Y) – вероятность появления предложения Y в языке, вычисляется с помощью модели языка (Language Model)

P(X|Y) – это обратный перевод, вероятность того, что конечное предложение Y можно перевести с помощью исходного предложения X, вычисляется с помощью модели перевода (Translation model)

Модель языка. Цепи Маркова

Способ оценивать вероятность всех теоретически возможных предложений языка

$P(x1,x2,\ldots,xn)=$

$$P(x1)*P(x2|x1)*P(x3|x1,x2)\ldots*P(xn|x1,x2,\ldots xn\text{-}1)$$

Цепи Маркова – статистические модели, позволяющие упростить вычисление условной вероятности.

Первого порядка

$P(x1,x2,\ldots,xn)= P(x1)*P(x2|x1)*P(x2|x3)\ldots*P(xn|xn\text{-}1)$

Второго порядка

$P(x1,x2,\ldots,xn)=$

$$P(x1)*P(x2|x1)*P(x2|x3)\ *P(x3|x1,x2)\ldots*\ \ldots*P(xn|xn\text{-}2,xn\text{-}1)$$

Триграммы: $q(c|a,b)$

Оценка максимального правдоподобия

maximum-likelihood estimates

q(c|a,b) – вероятность появления слова *c* при условии появления предшествующих слов *a* и *b*

Как ее оценить?

$$q_{ml}(c|a,b) = \frac{K(a,b,c)}{K(a,b)}$$

K(a,b) – количество биграмм ab в корпусе

K(a,b,c) – количество триграмм abc в корпусе

Оценка максимального правдоподобия: примеры

Примеры

a= кандидат

b= филологических $q(c|a,b)=1$

c= наук

Мама мыла раму. a=мама

Мама мыла машину. b=мыла $q(c|a,b)=1/3$

Мама мыла белую раму. c= раму

Оценка максимального правдоподобия: недостатки

- Большинство потенциально возможных триграмм в корпусе не встретится, q=0
- Если в корпусе нет какой-либо биграммы, знаменатель в формуле будет =0.

Преодоление этих недостатков – методы сглаживания (smoothed estimation methods) - способ оценки условной вероятности, который всегда дает ненулевые значения.

Методы сглаживания: линейная интерполяция

$$q_{ml}(c|b) = \frac{K(b,c)}{K(b)} \; ; \; q_{ml}(c) = \frac{K(c)}{K(\,)}$$

Вероятность появления слова с после слова b

Вероятность появления слова с без учета контекста

$$\lambda_1 + \lambda_2 + \lambda_3 = 1$$

$$q(c|a,b) = \lambda_1 * q_{ml}(c|a,b) + \lambda_2 * q_{ml}(c|b) + \lambda_3 * q_{ml}(c)$$

Методы дисконтирования

Резервирование вероятности для сочетаний, которых нет в обучающем корпусе. От числа появления каждой n-граммы надо отнять небольшое число β (часто β=0.5)

$$q_{ml}(c|b) = \frac{K(b,c) - \beta}{K(b)}$$

Сэкономленная вероятность = β * Kvo(c,v)

Kvo(c,v) – число различного вида биграмм со словом ***c***

Делится между всеми теоретически возможными биграммами (c,y), которые не встретились в обучающем корпусе

Модель статистического перевода: задачи

Позволяет оценить вероятность того, что одно предложение является переводом другого: что заданное предложение X является исходным для предложения Y

X = This cat is nice

Y1= Этот кошка хорош	P(Y1|X)
Y2 = Эта этот кот	P(Y2|X)
Y3 = Эта кошка хороша	P(Y3|X)
Y4 = Эта кошка есть красивый	P(Y4|X)
Y5 = Этот кот красивый	P(Y5|X)
Y6= Вася – дурак	P(Y6|X)

Модель статистического перевода: выравнивание

Выравнивание (alignment):

- по документам
- по абзацам
- по предложениям (построение параллельного корпуса предложений); предположение о монотонности, одинаковом порядке предложений
- по словам: массив выравнивания

Выравнивание по предложениям. Алгоритмы:

- без использования лексической информации: длинные фрагменты переводятся в длинные фрагменты
- использование лексики: небольших двуязычных словарей
- использование дополнительной информации (даты, числа)

Модель статистического перевода: виды

Модель перевода может быть основана на:

- словах

  Word-based translation — WBT

- словосочетаниях

  Phrase-based translation — PBT

- синтаксисе

  Syntax-based translation — SBT

- …

Перевод, основанный на словах: схема

$a: 1 \to 1, 2 \to 2, 3 \to 0, 4 \to 5, 5 \to 3, 6 \to 4$

- в разных языках разный порядок слов
- некоторые слова могут не иметь перевода
- некоторые слова переводятся несколькими словами

Перевод, основанный на словах: данные

Данные для вычислений на основе параллельного размеченного корпуса:

- набор исходных предложений (каждое из m_p слов)
- набор конечных предложений, каждое состоит из n_p слов)
- массив выравнивания

Необходимые вычисления для каждой тройки (x,y,a)

- сколько раз в исходном корпусе встретилось каждое слово из предложения X

- сколько раз это слово было переведено словом из Y

- сколько раз слово номер i в X связано со словом номер j в предложении Y

- сколько раз позиция i встречалась в предложении X при определенных длинах предложений X и Y

Вычисляется оценка максимального правдоподобия

EM-алгоритм – алгоритм увеличения «ожидаемого правдоподобия» (Expectation-maximization algorithm)

$$c(e|f;\mathbf{e},\mathbf{f}) = \frac{t(e|f)}{\sum_{i=0}^{n} t(e|f_i)} \sum_{j=1}^{k} \delta(e,e_j) \sum_{i=0}^{n} \delta(f,f_i)$$

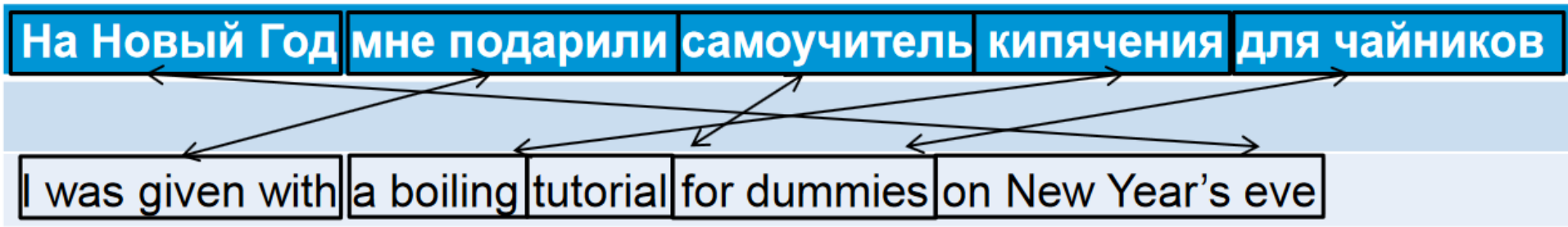
Способы декодирования

- Полный перебор
- Поиск по первому наилучшему совпадению
- Жадный инкрементный поиск
- Сведение к обобщенной задаче коммивояжера
- Метод Лина-Кернигана
- Генетические алгоритмы
- Алгоритмы муравьиной оптимизации

Идеи:

-«Подбор» продолжения предложения
- Выбор лучшей на данный момент альтернативы
- Изменение предложения для улучшения перевода

Перевод, основанный на словосочетаниях



- Переводить по несколько слов сразу
- Сохраняется контекст, а, следовательно, смысл
- Изменение порядка слов все равно необходимо

Алгоритм

1. Построить выравнивания по словам
2. Собрать все фразы, согласующиеся с выравниванием по словам, т.е. фразы должны включать все точки выравнивания содержащихся в них слов в обоих языках
3. Посчитать вероятности

Перевод на основе синтаксических деревьев

Позволяют включить синтаксическую информацию:

- Перестановка слов на основе синтаксиса
- Более правильное использование вспомогательных слов
- Грамматически более верный перевод

Особенности:

- Эффективность сравнима с моделями на словосочетаниях
- Нужны хорошие синтаксические парсеры
- Вычислительно сложные

Статистический перевод: особенности

- **Достоинства** : быстрая настройка основных модулей по корпусу (машинное обучение)
- **Недостатки**:
  - дефицит параллельных корпусов текстов
  - нестабильность перевода, т.к. зависимость
    — от статистики: *Лев Толстой – Lion Thick*
    — от грамматических ошибок и опечаток в текстах

Качественный перевод возможен только для фраз, которые целиком помещаются в n-граммную модель

Гибридный перевод

Time flies as arrow

Время летит как стрела *Мухи времени как стрела*

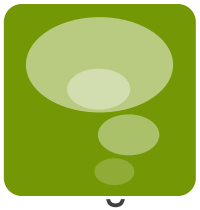
Направления:

- Интеграция статистического модуля в перевод, основанный на правилах (перевод по правилам, затем выбор нужного варианта по статистике)

- Интеграция правил в статистическую модель (предобработка исходного текста с помощью правил)

Пример: система PROMT – параллельный корпус исправленных ошибок для работы модуля синтаксического постредактирования

Нейронный МП



- Целью НМП является создание полностью обучаемой модели, каждый компонент которой настраивается на основе обучающих корпусов, чтобы максимизировать качество перевода.
- Полностью обучаемая НМП-модель генерирует как можно более естественное представление целевого предложения.
- Используются рекурентные нейронные сети
- Возможность учитывать не только типичное значение фразы, но и ее контекст

Google? Yandex?

Оценки качества МП

- Ручная (не менее 4 человек)
  - по параметрам (оценка в баллах):
    полнота (adequacy) – точность перевода,
    гладкость (fluency) – правильность фразы
  - выбор лучшего варианта, рейтинг вариантов
  - оценка ресурсов, необходимых для постредактирования (технология post-editing distance)
- Автоматическая – сравнение текста с эталоном на основе метрики

Метрики качества МП

- Одна из первых метрик –

  WER (Word Error Rate) =
  расстояние Левенштейна между двумя переводами, деленное на длину образцового перевода

- Самая популярная –

  мера *BLUE* (*Bilingual Evaluation Understudy*) ,
  вычислительно достаточно проста:
  = отношение количества N-грамм из образцового перевода, найденных в переводе системы,
  к количеству всех N-грамм перевода, N= 4

BLUE: расчет

$$BLEU_4 = min\left(1, \frac{\text{длина перевода}}{\text{длина эталонного перевода}}\right) \prod_{i=1}^{4} P_i,$$

$$\text{где } P_i = \frac{\text{Кол–во совпавших } i\text{–грамм}}{\text{Кол–во } i\text{–грамм}}$$

	Как упоительны в России вечера
Эталон	So delightful are evenings in Russia
Перевод G	As a delightful evening in Russia
Перевод P	As evenings are delightful in Russia
Перевод Y	As упоительны in Russia PM

$$P_1 = \frac{5}{6}, P_2 = \frac{1}{5}, P_3 = 0$$

$$BLEU_2 = \frac{6}{6} \cdot \frac{5}{6} \cdot \frac{1}{5} = 16\%$$

$$BLEU_4 = 0$$

BLUE: особенности

Достоинства

- Есть положительная корреляция с оценкой перевода человеком (не бесспорно)

Недостатки

- Не учитывает важность слов
- Не учитывает грамматическую корректность, т.к. работает на уровне фраз из 4-х слов
- Значение BLEU не имеет смысла, имеет смысл только разница значений
- BLEU человеческих переводов не сильно выше автоматических, хотя сам перевод лучше

Наиболее известные и современные системы МП

Translate.Ru (Promt)
Google Translate
Яндекс. Переводчик
Systran
OpenLogos
Babylon
ABBYY Compreno

Voila
Applied Language Solutions Online
Reverso
Trident
InterTran
Linguatec

Windows Live Translator Bing
FreeTranslation
SDL Translate
ImTranslator
WorldLingo
Babel Fish
Pragma 5.x

Система Systran

Systransoft.com

США, затем Франция, затем Корея

Самая старая из успешных и самая успешная из старых систем МП

Основана на правилах, 52 языка

Начало разработки 1968 г. (для военно-воздушных сил США), создан переводчик с русского на английский (холодная война)

1973 г. – переводчик с английского на русский (космический проект Союз-Аполлон)

На основе Systran до 2007 г перевод в Yahoo!,. Google

Системы
Logos и OpenLogos

США и Германия

1970 г. Для войны во Вьетнаме; англо-вьетнамский

Коммерческая версия программы сейчас разрабатывается компанией Group Business Software AG для операционной системы Microsoft Windows

2005 г. – OpenLogos – бесплатная версия

Немецкий, английский ←→французский, итальянский, испанский, португальский

На основе правил

Система Babylon

Израиль, 1997 г.

http://perevodchik.babylon-software.com

2011г. – лидер по числу бесплатных скачиваний (в книге рекордов Гиннеса)

Сайт в числе 100 наиболее посещаемых сайтов мира

Более чем 1600 словарей и глоссариев из самых разнообразных сфер и областей на более чем 75 языках мира.

Google Translate

США

Вначале на основе Systran

2004 г. – разработка собственных статистических методов: арабский-английский, русский – китайский

2007 г. – все имеющиеся языковые пары переведены на статистические методы; 73 языка, например, с телугу на хмонг)

В настоящее время Google Translate пользуются более 500 млн человек ежедневно, которые переводят свыше 100 млрд слов на 103 языка

Сообщество сервиса насчитывает свыше 3,5 млн активных участников, которые внесли порядка 90 млн корректировок.

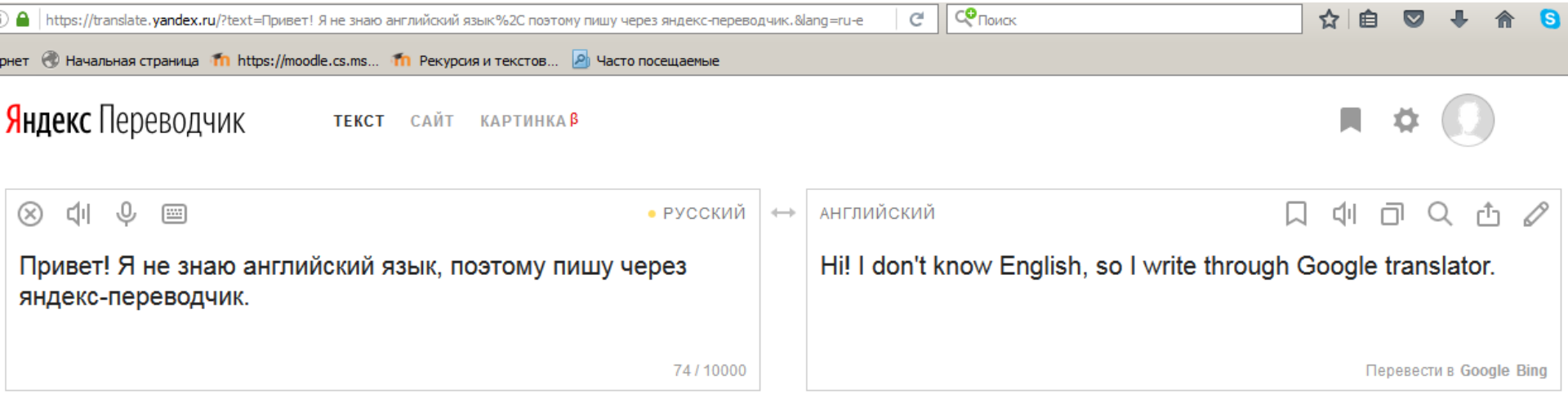
Яндекс. Переводчик

Россия

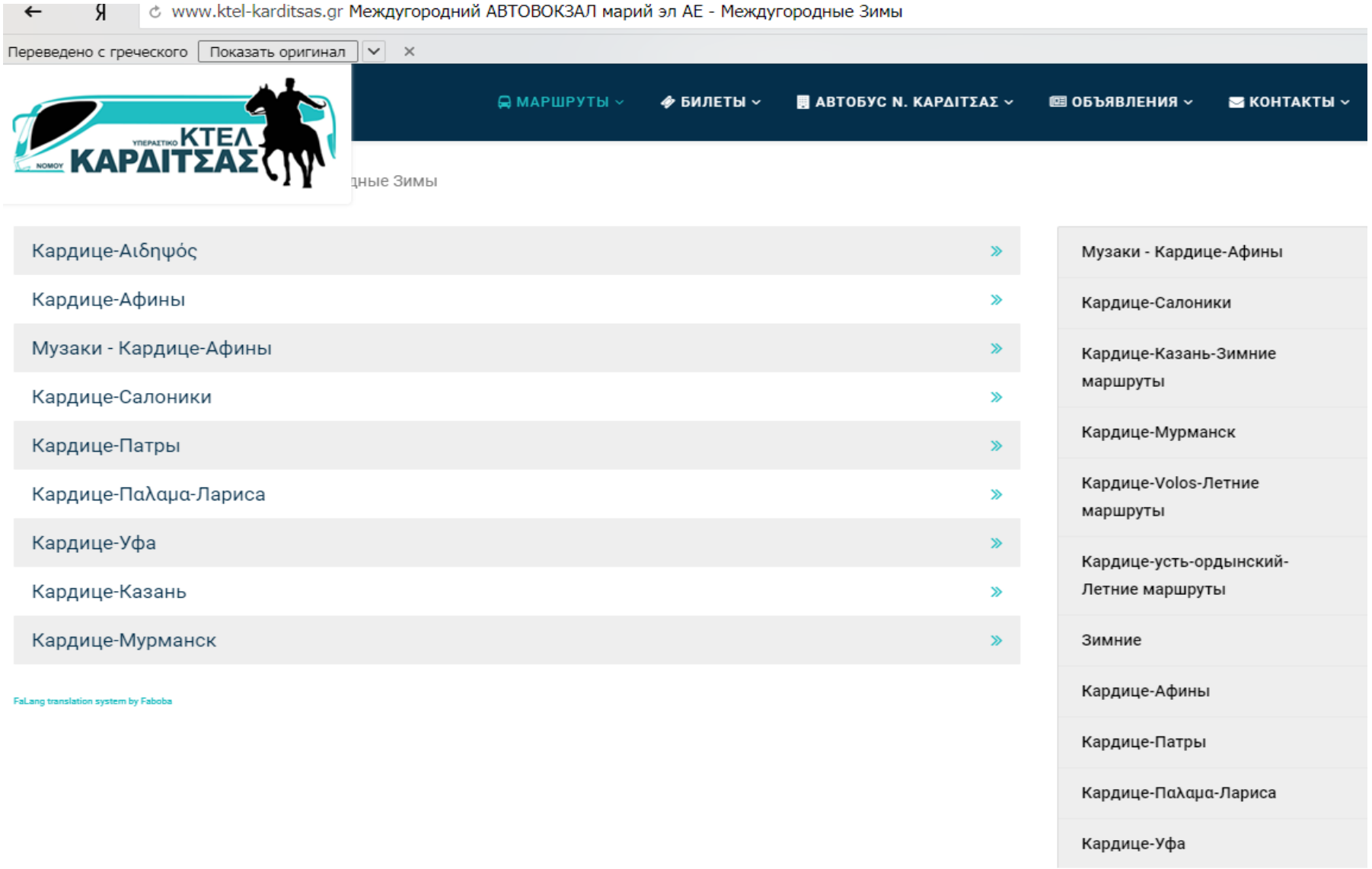
Март 2011: перевод текстов и веб-страниц русский-английский-украинский

Сейчас – перевод для 67 языков

Статистические методы



Яндекс. Переводчик: Пример



Система ABBYY Compreno

Россия

abbyy.com

2011 – грант от Сколково для создания новой технологии синтаксического и семантического анализа текста

«**ABBYY Compreno** - это уникальная технология анализа и понимания текстов на естественном языке. В отличие от систем, основанных на статистике и правилах, ABBYY Compreno выполняет полный семантико-синтаксический анализ текста, создает его универсальное представление, извлекает сущности, события и связи между ними.

Решение разных задач прикладной лингвистики, в том числе автоматический перевод.

Система PROMT

Россия, 1991 г. (с 1992 по 1998 STYLUS)

64 языковые пары (34 пары без использования русского языка)

Вначале – на основе правил, с 2010 г. – статистические и гибридные технологии

2011, 6-й ежегодный семинар по статистическому МП: перевод с английского на немецкий – первое место, перевод с английского на испанский – 4-е место из 15

2013-14 гг., 8-й и 9-й семинар : 1-е место по переводу с английского на русский

Семинар по статистическому машинному переводу

http://statmt.org/

Ежегодный семинар

Под эгидой Ассоциации компьютерной лингвистики (ACL), основанной в 1962 г. (http://www.aclweb.org/)

Используются тексты стенограммы заседаний и тексты документов Европарламента, доступные как раз для основных европейских языков, а также новостные тексты.

Экспертиза – использование краудсорсинга (привлечение большого количества экспертов)

Сравнение систем

Имеет смысл только сравнение систем между собой на определенных текстах

	Promt	***Google***	***Systran***	***Windows***
Русский - английский	*8*	*8*	*8*	*8*
Английский - русский	*4*	*5*	*2*	*5*
Немецкий - русский	*4*	*5*	*-*	*5*
Французский - русский	*4*	*4*	*-*	*4*
Китайский - русский	*-*	*2*	*-*	*-*
Арабский - русский	*-*	*4*	*-*	*3*

Сравнение текстов переводов

- Substantial advances to promote democracy meant that this country was no longer a threat to America's national security.

Варианты:

- Достигнуты значительные успехи по продвижению демократии означает, что эта страна больше не представляет угрозы для национальной безопасности Америки (Yandex)
- Существенные достижения, чтобы способствовать демократии означали, что эта страна больше не была угроза национальной безопасности Америки. (Promt)
- Существенный прогресс по продвижению демократии не означает, что эта страна больше не представляет угрозы для национальной безопасности Америки (Google)

Заключение

Реальность – Маленький Джон по-прежнему ищет свою коробку в ручке (ну, в крайнем случае, в загоне)

Перспективы

- **2020** "Обучение детей чтению и письму - бесполезная трата времени", заявил Йео Киа Вей, министр образования Сингапура и упразднил изучение этих предметов в школе. "Дети должны быть избавлены от тягостного труда по распознаванию крошечных значков на бумаге или на экране. Пусть этим занимаются машины".
- **2043** Строительство Вавилонской башни завершено в Ираке (бывшей Вавилонии) после 4000-летнего перерыва, благодаря Универсальному языку компании NEC Technologies.

Задание

Перевести один и тот же небольшой текст разными переводчиками. Оценить качество перевода одним из способов (указать, каким).

Подобрать текст, который будет хорошо переводиться, и текст, который будет переводиться плохо. Почему?

Ссылки

https://translate.yandex.ru/

https://translate.google.com/

http://www.translate.ru/#!/auto/