Донейросетевые методы машинного перевода

Глазкова Екатерина, БПМИ152

НИУ ВШЭ, 04.12.2017

Содержание

- Машинный и автоматизированный перевод
- Виды МП
  - Перевод на базе правил
  - Статистический МП
    - Модель перевода
    - Модель языка
    - Декодер
- Донейросетевой Яндекс Переводчик
- Метрики качества перевода
  - Word error rate
  - BLEU

Машинный и автоматизированный перевод

Машинный перевод (Автоматический, Machine Translation, MT)

- Процесс перевода текстов специальной компьютерной программой
- Направление научных исследований

Автоматизированный перевод (Machine-Aided Translation, MAT)

- Процесс перевода осуществляется человеком, компьютер помогает перевести текст за меньшее время и с лучшим качеством
- Проверка правописания, компьютерные словари, индексаторы

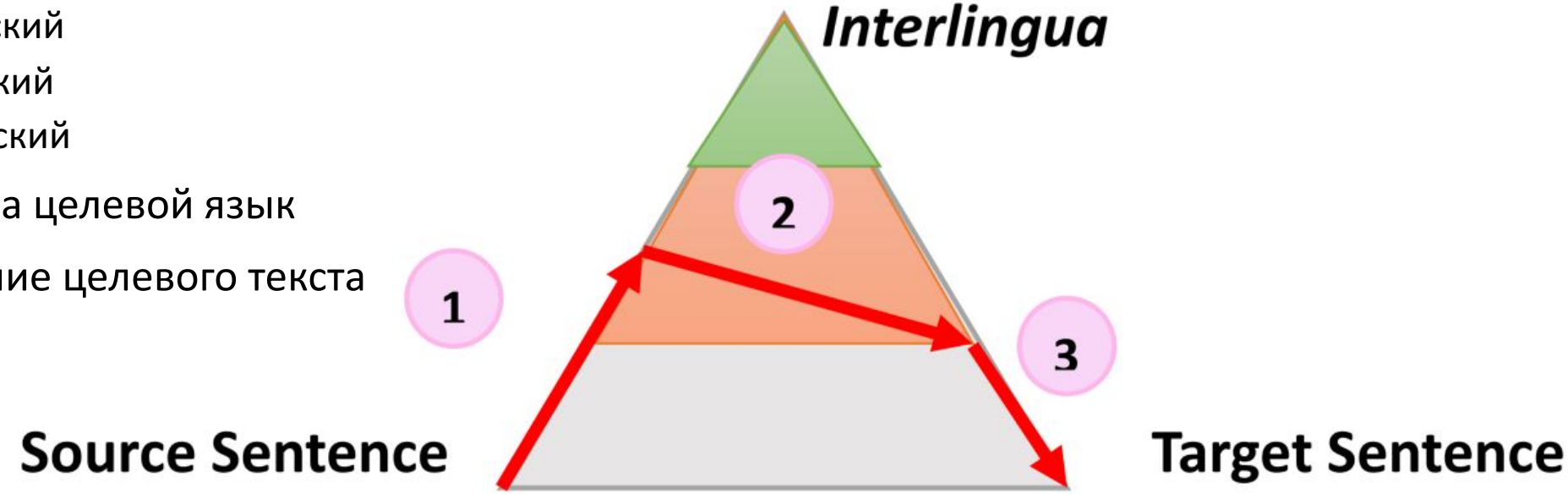
Виды машинного перевода

- На основе правил (Rule-based machine translation – RBMT)
- Статистический (Statistical machine translation – SMT)
- Перевод на основе примеров (Example-Based – EBMT)
- Нейронный (Neural Machine Translation – NMT)
- Гибридный (Hybrid machine translation)

Треугольник Вокуа

Этапы перевода:

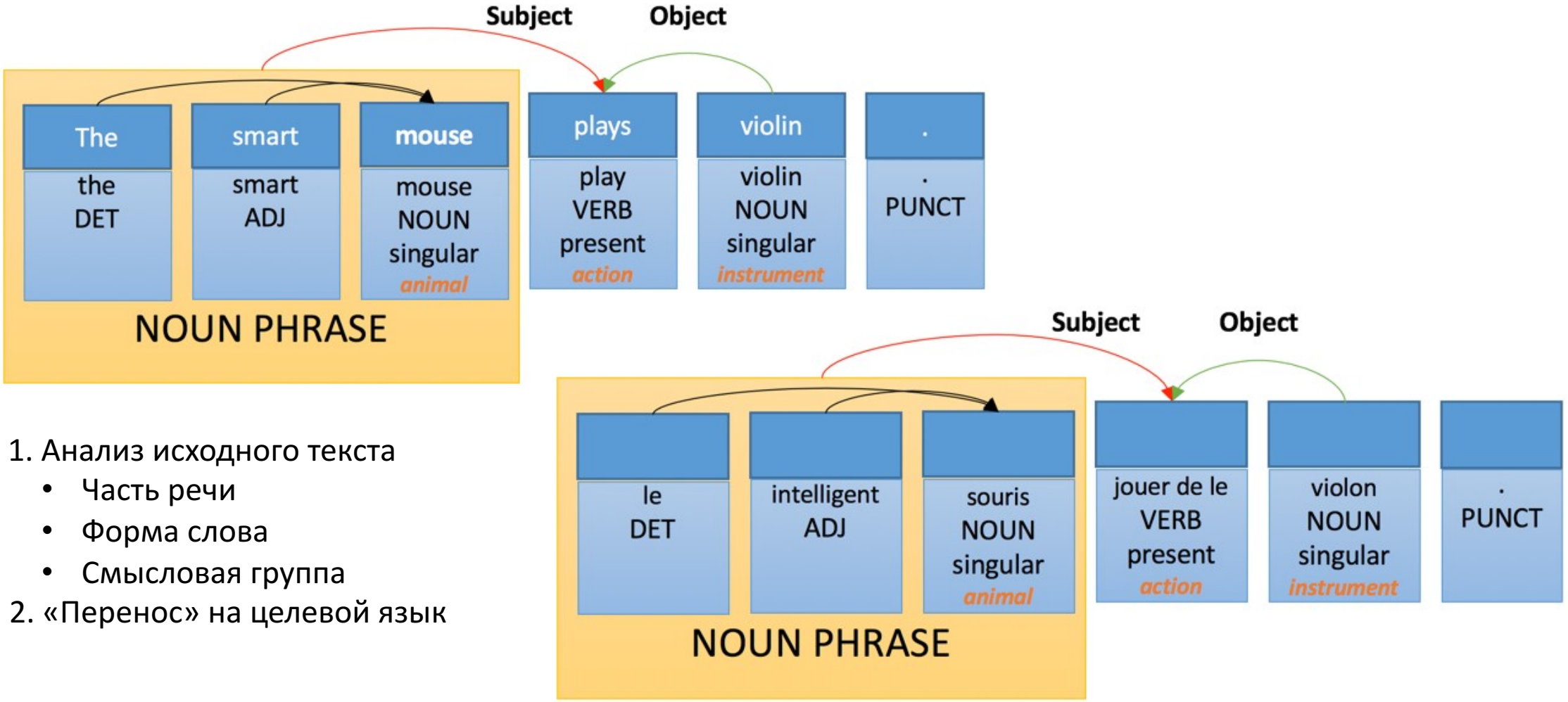
1. Анализ исходного текста
   - Морфологический
   - Синтаксический
   - Семантический
   - Прагматический
2. «Перенос» на целевой язык
3. Формирование целевого текста



«Классический» треугольник Вокуа

http://blog.systransoft.com/how-does-neural-machine-translation-work/

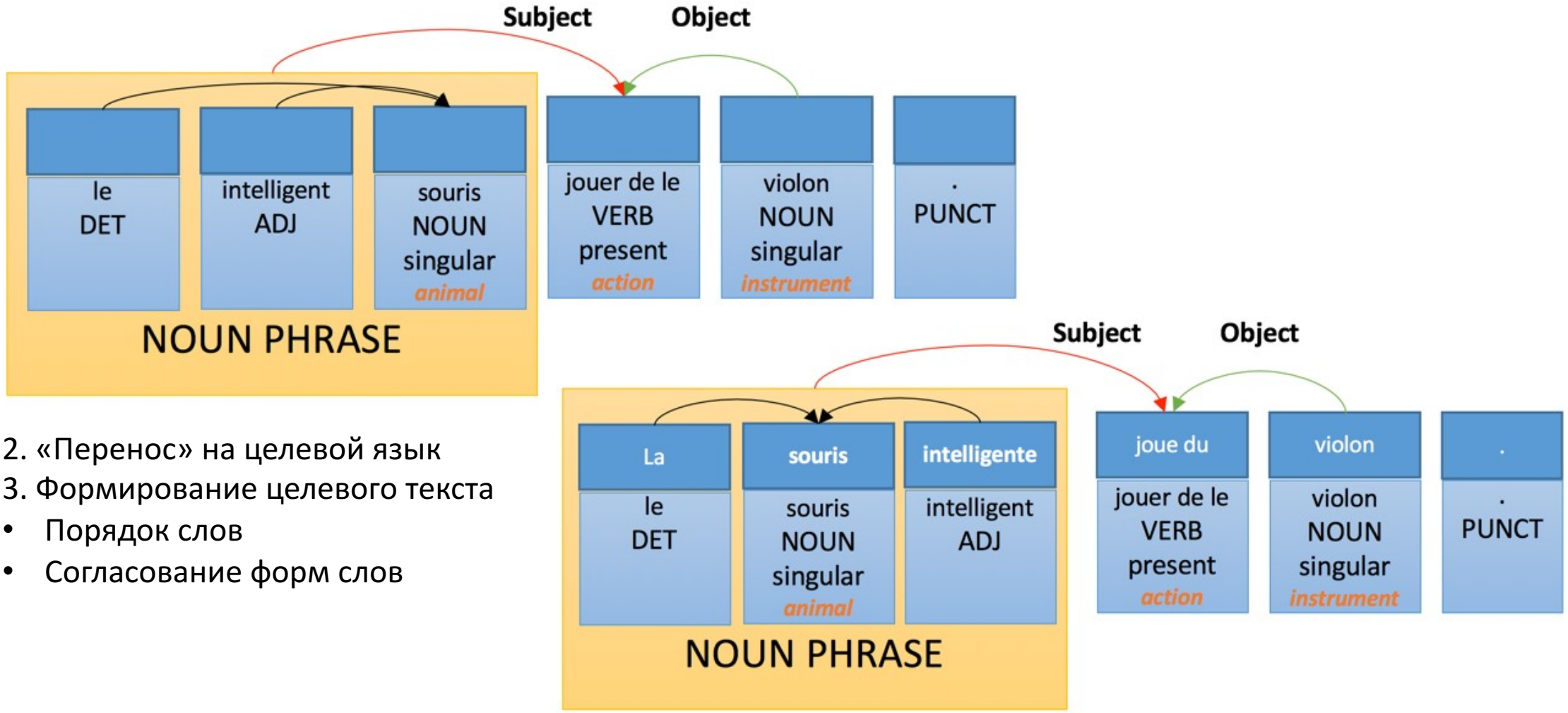
Перевод на базе правил



1. Анализ исходного текста
   - Часть речи
   - Форма слова
   - Смысловая группа
2. «Перенос» на целевой язык

http://blog.systransoft.com/how-does-neural-machine-translation-work/

Перевод на базе правил (продолжение)

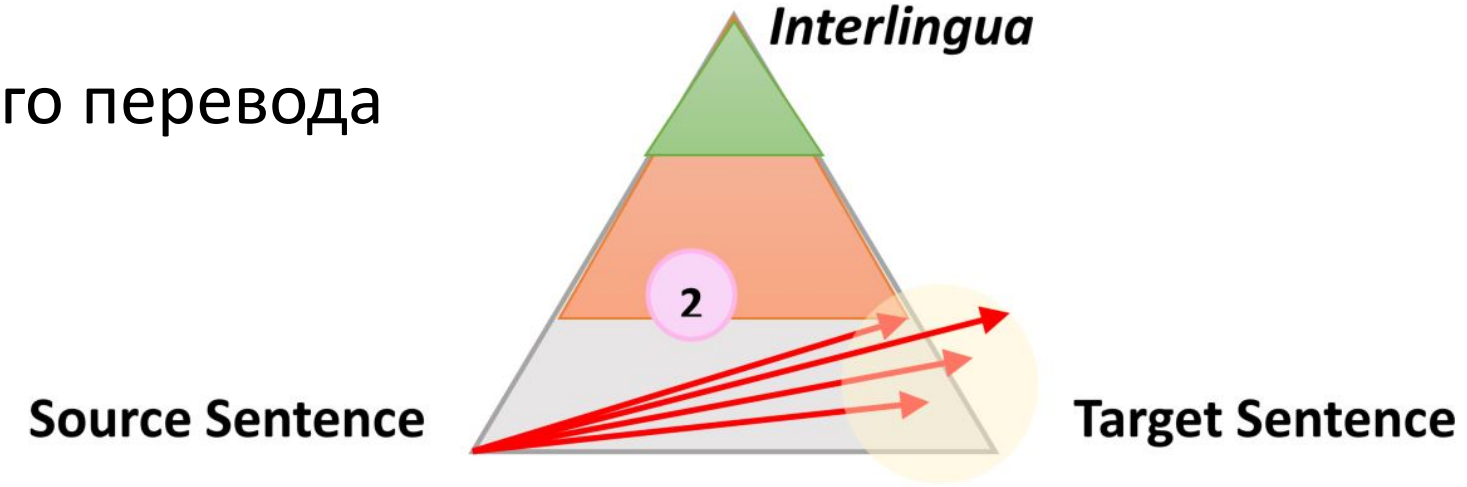


2. «Перенос» на целевой язык
3. Формирование целевого текста
   - Порядок слов
   - Согласование форм слов

http://blog.systransoft.com/how-does-neural-machine-translation-work

Статистический машинный перевод

- Разделение данных на блоки (слова/фразы)
- Перевод блока несколькими способами
- Выбор оптимального перевода

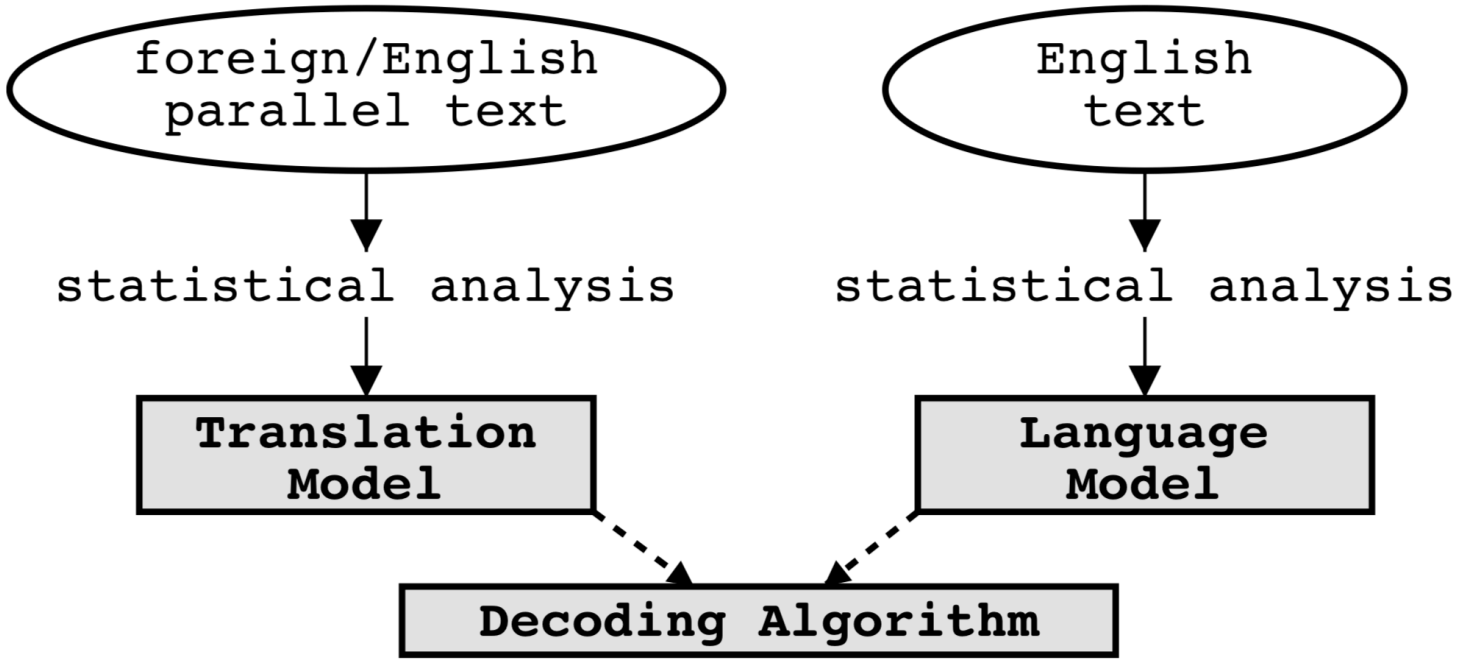


http://blog.systransoft.com/how-does-neural-machine-translation-work

Статистический машинный перевод

Состоит из 3 частей:

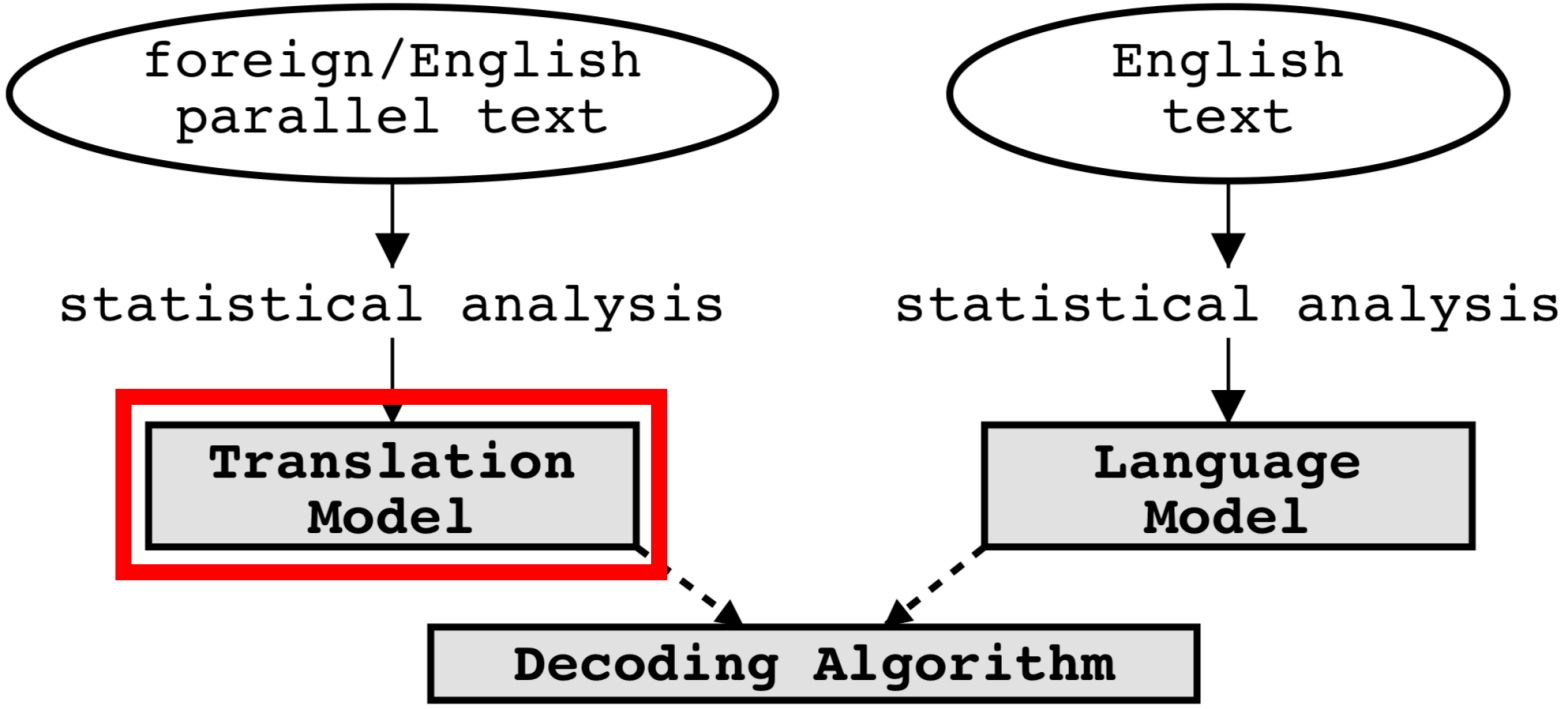
- модель перевода
- модель языка
- декодер



Модель статистического МП на английский язык

http://www.mt-archive.info/MTS-2007-Koehn-3.pdf

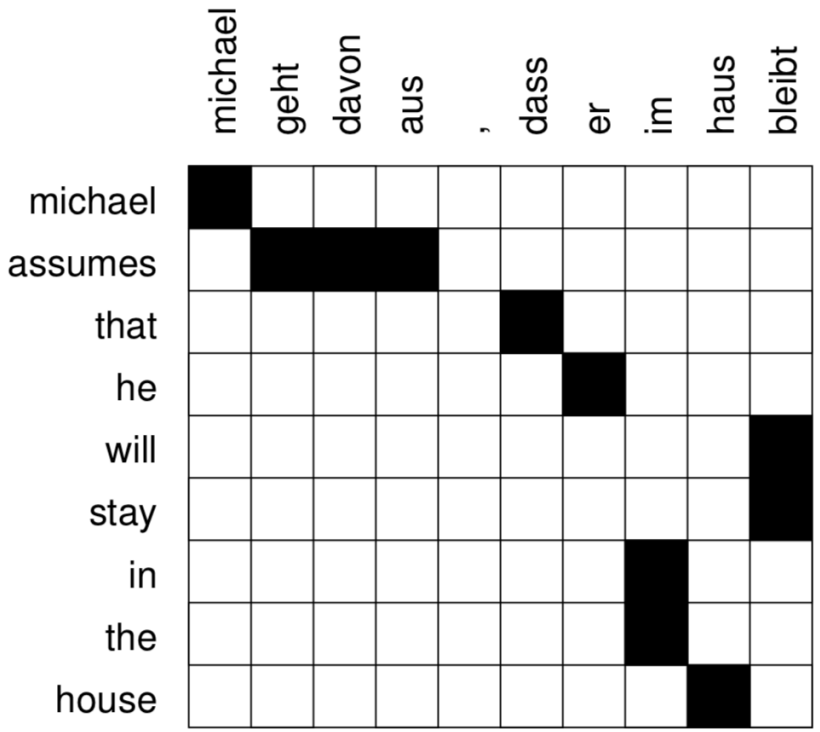
Модель перевода



Модель перевода

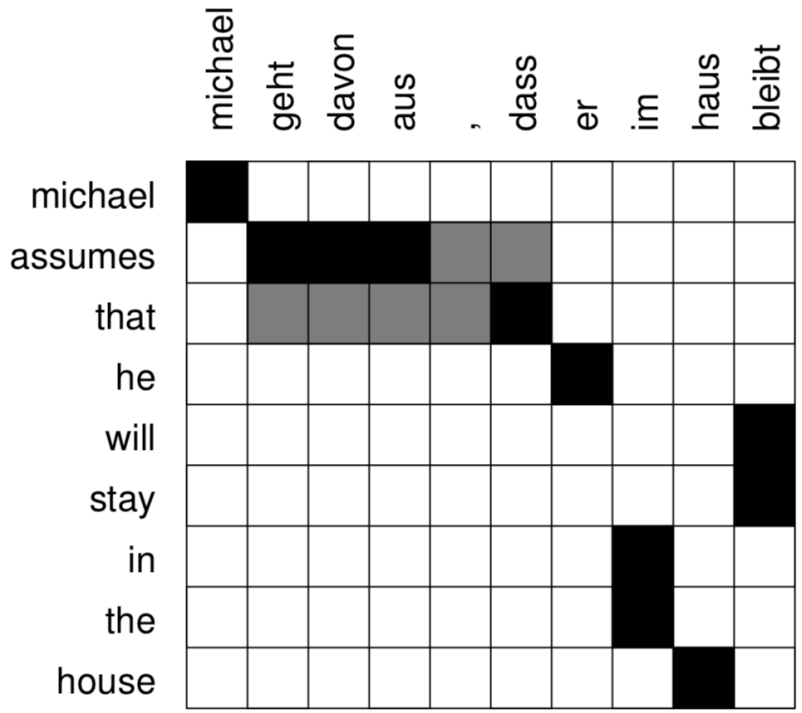
- Таблица всевозможных переводов с указанием вероятности каждого
- Составляется для каждой пары языков
- Строится на параллельных текстах

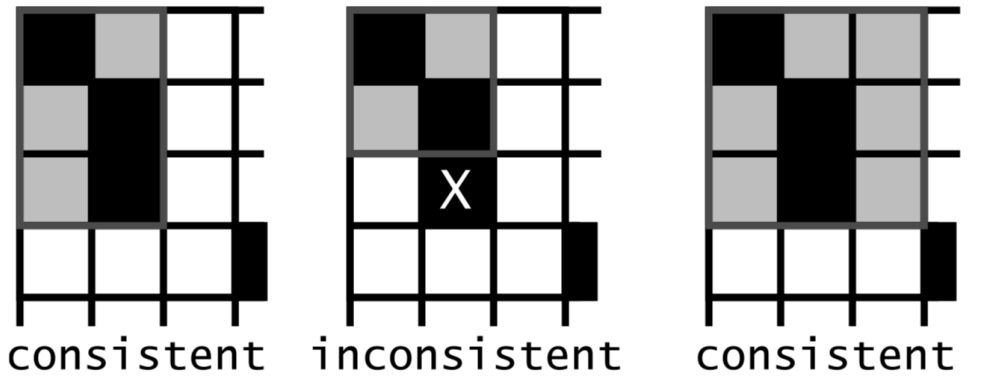
Построение модели перевода
Выделение расположения слов



http://www.statmt.org/book/slides/05-phrase-based-models.pdf
www.cs.cornell.edu/courses/cs5740/2017sp/lectures/08-alignments.pdf

Построение модели перевода
Извлечение фраз





ok

violated
one alignment
point outside

ok
unaligned
word is fine

http://www.statmt.org/book/slides/05-phrase-based-models.pdf

Построение модели перевода
Оценивание вероятности фраз

$$\phi(\overline{g_j}|\overline{e}) = \frac{count(\overline{e}, \overline{g_j})}{\sum_{\overline{g_i}} count(\overline{e}, \overline{g_i})}$$

$\overline{e}$ - фраза на языке оригинала

$\overline{g_i}$ - вариант фразы на языке таргета

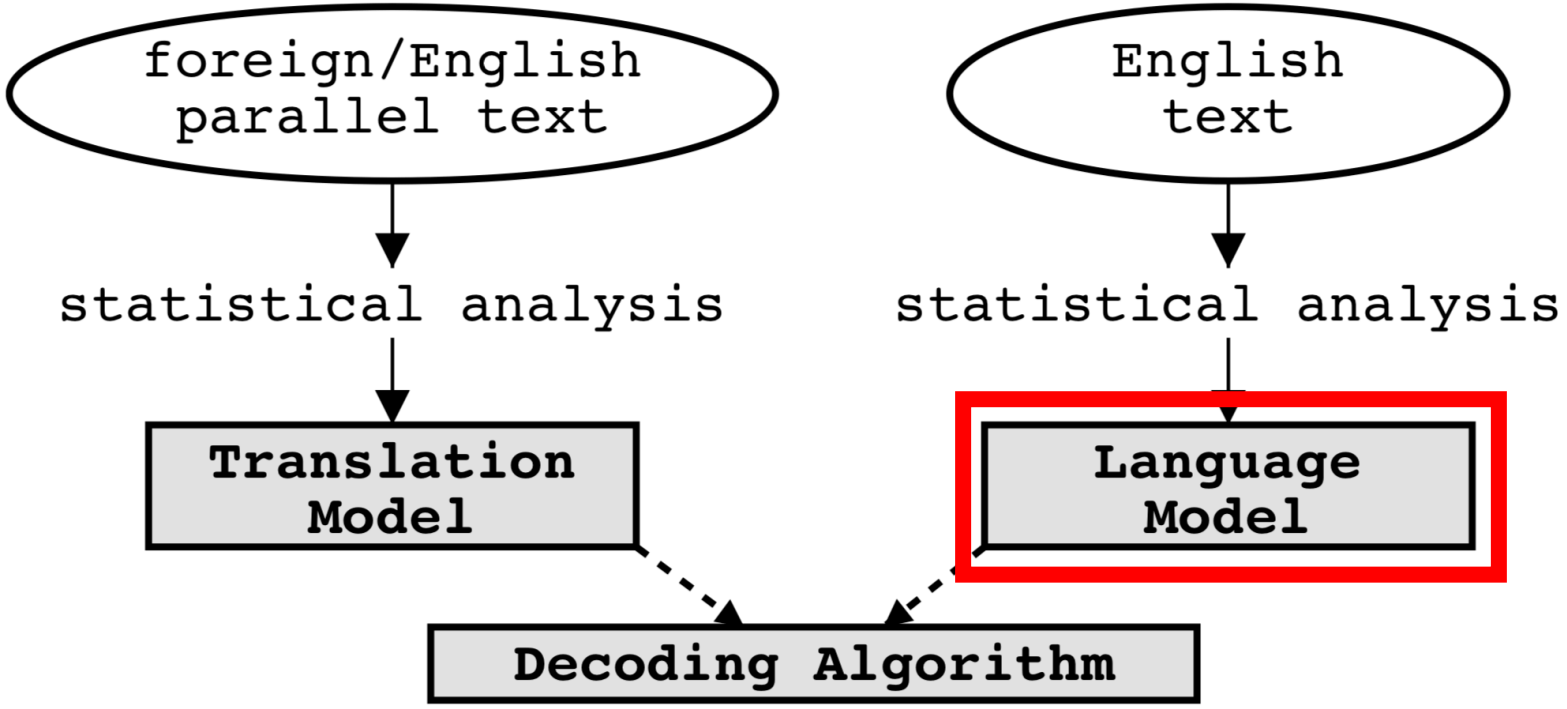
Пример модели перевода

English	$\phi(\bar{e}\|\bar{f})$	English	$\phi(\bar{e}\|\bar{f})$
the proposal	0.6227	the suggestions	0.0114
's proposal	0.1068	the proposed	0.0114
a proposal	0.0341	the motion	0.0091
the idea	0.0250	the idea of	0.0091
this proposal	0.0227	the proposal ,	0.0068
proposal	0.0205	its proposal	0.0068
of the proposal	0.0159	it	0.0068
the proposals	0.0159	...	...

Таблица перевода для немецкого слова den Vorschlag

http://www.statmt.org/book/slides/05-phrase-based-models.pdf

Модель языка



Модель языка

- Частота использования фраз в языке
- Составляется для одного языка

- Учет порядка слов

$$p_{LM}(\text{" the house is small "}) > p_{LM}(\text{" small the is house "})$$

- Учет особенностей синонимов

$$p_{LM}(\text{" I am going home "}) > p_{LM}(\text{" I am going house "})$$

Построение модели языка

Пусть W = w_1, w_2,…, w_n – строка

Тогда

$$p(W) = p(w_1, w_2, \dots w_n) = p(w_1)p(w_2|w_1)p(w_3|w_1, w_2) \dots p(w_n|w_1, w_2, \dots, w_{n-1})$$

Для 2-Gram:

$$p(W) = p(w_1, w_2, \dots w_n) \backsimeq p(w_1)p(w_2|w_1)p(w_3|w_2) \dots p(w_n|w_{n-1})$$

$$p(w_2|w_1) = \frac{count(w_1, w_2)}{count(w_1)}$$

Пример модели языка

the green (total: 1748)

word	c.	prob.
paper	801	0.458
group	640	0.367
light	110	0.063
party	27	0.015
ecu	21	0.012

the red (total: 225)

word	c.	prob.
cross	123	0.547
tape	31	0.138
army	9	0.040
card	7	0.031
,	5	0.022

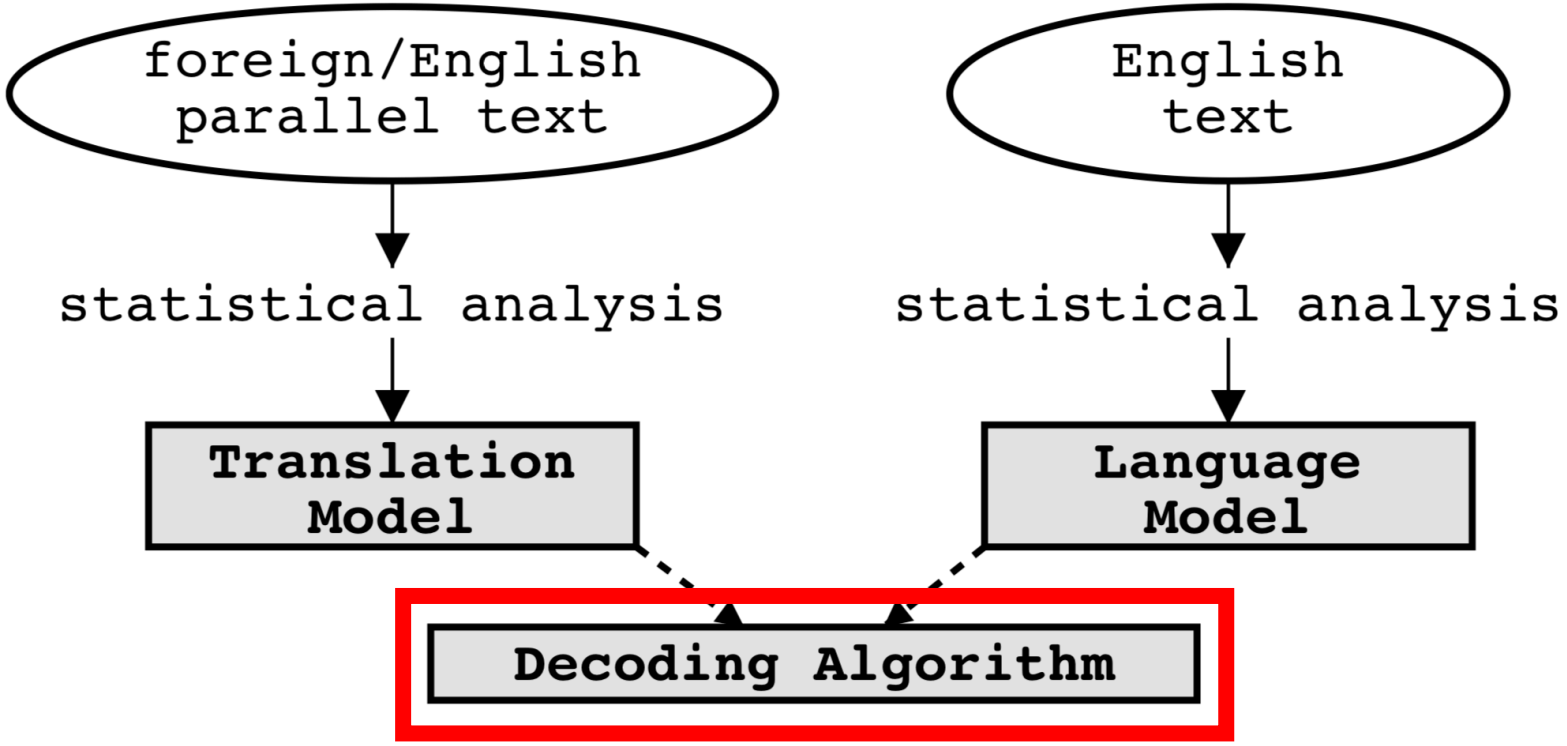
the blue (total: 54)

word	c.	prob.
box	16	0.296
.	6	0.111
flag	6	0.111
,	3	0.056
angel	3	0.056

Модель языка при использовании 3-Gram

http://www.statmt.org/book/slides/07-language-models.pdf

Декодер


foreign/English parallel text
English text
statistical analysis
statistical analysis
Translation Model
Language Model
Decoding Algorithm

Декодер

Решение задачи

$$e_{best} = argmax_e p(e|g) = argmax_e p(g|e) p_{LM}(e)$$

g – исходный текст

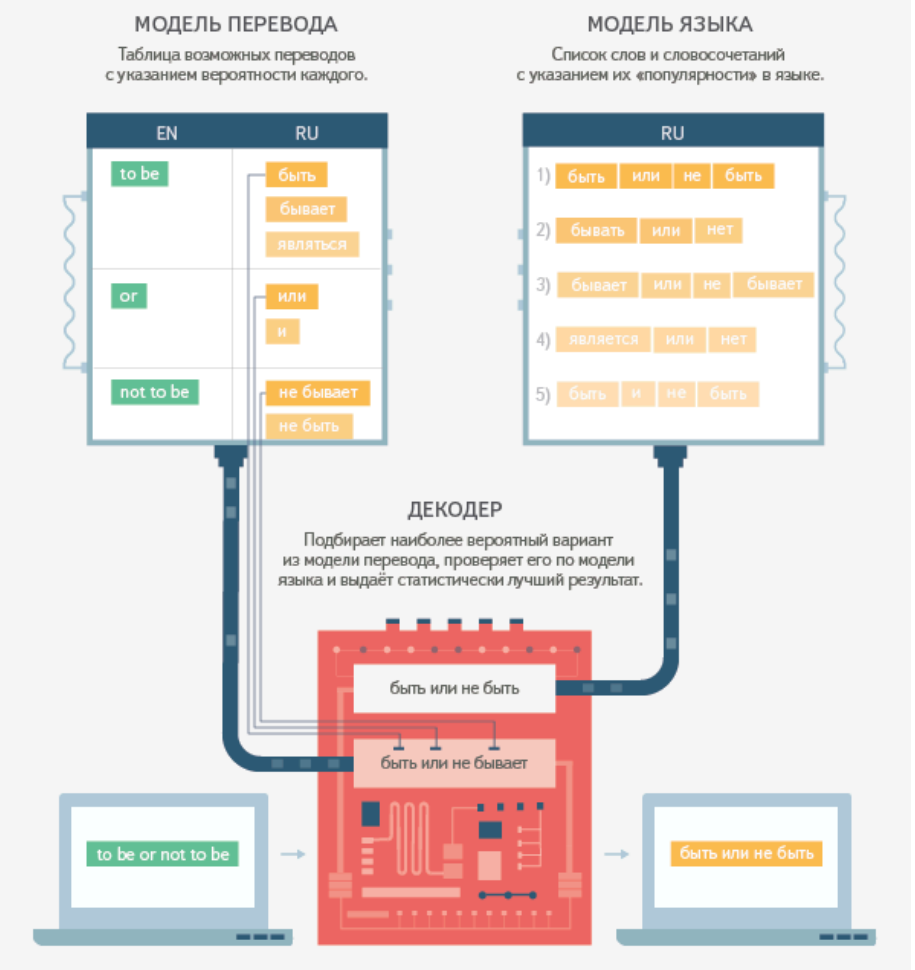
e – переведенный текст

p(e|g) – модель перевода

p_{LM}(e) – модель языка

Донейросетевой Яндекс Переводчик

- 16 марта 2011 - запуск сервиса статистического МП Яндекс.Перевод
- 14 сентября 2017 – запуск гибридного переводчика
- Алгоритм:
  - Подбор всех вариантов перевода
  - Сортировка вариантов перевода
  - Оценка частоты употребления для всех вариантов с помощью модели языка
  - Выбор оптимального сочетания вероятности перевода и частоты употребления



https://yandex.ru/company/technologies/translation/

Измерение качества перевода

Трудность: нет единственно правильного перевода

Виды метрик:

- Оценка вручную
  - смысл (Adequacy)
  - гладкость речи (Fluency)
- Автоматические метрики:
  - Word Error Rate
  - BLEU
- Task-based
  - Round-Trip Translation (source-> target -> source)
  - Объем постредактирования

Word Error Rate

- Редакционное расстояние Левенштейна, но для слов

$$WER = \frac{substitutions + insertions + deletions}{reference\_length}$$

BLEU (bilingual evaluation understudy)

- Коррелирует с оцениванием перевода человеком
- Самая популярная и часто реализуемая метрика

$$BLEU = min\left(1, \frac{output\_length}{reference\_length}\right)\left(\prod_{i=1}^{n} precision_i\right)^{\frac{1}{n}}$$

n – максимальная длина учитываемых сочетаний слов

Пример BLEU

SYSTEM A: Israeli officials (2-GRAM MATCH) responsibility of airport (1-GRAM MATCH) safety

REFERENCE: Israeli officials are responsible for airport security

SYSTEM B: airport security (2-GRAM MATCH) Israeli officials are responsible (4-GRAM MATCH)

Metric	System A	System B
precision (1gram)	3/6	6/6
precision (2gram)	1/5	4/5
precision (3gram)	0/4	2/4
precision (4gram)	0/3	1/3
brevity penalty	6/7	6/7
BLEU	0%	52%

http://www.statmt.org/book/slides/07-language-models.pdf

Источники информации

- Philipp Koehn. Statistical Machine Translation. Cambridge University Press, 2009. 488 - http://www.statmt.org/book/
- Philipp Koehn, Franz Josef Och, Daniel Marcu. Statistical Phrase-Based Translation, 2003, University of Southern California - https://www.isi.edu/~marcu/papers/phrases-hlt2003.pdf
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a Method for Automatic Evaluation of Machine Translation. IBM T. J. Watson Research Center - http://www.aclweb.org/anthology/P02-1040.pdf
- Harold Somers. Round-Trip Translation: What Is It Good For? Manchester University, 2005 - http://www.mt-archive.info/ALTW-2005-Somers.pdf

- Блог компании Systran – http://blog.systransoft.com/how-does-neural-machine-translation-work
- Блог компании Яндекс – https://yandex.ru/company/technologies/translation
- Википедия