

Донейросетевые методы машинного перевода

Жилкина Ксения

Факультет Компьютерных Наук, ПМИ
НИУ ВШЭ

7 декабря 2018

Оглавление

- Машинный перевод

 - История

- Перевод на основе правил

 - Системы дословного перевода

 - Трансферные системы

 - Интерлингвистические системы

 - Преимущества и недостатки RBMT

- Перевод на основе примеров

 - Близость предложений

 - Рекомбинация

 - Преимущества и недостатки EBMT

- Статистический перевод

 - Статистический перевод по словам

 - Статистический перевод по фразам

 - SMT на основе синтаксиса

 - Преимущества и недостатки SMT

- Донейросетевой МП: что сейчас и что дальше?

Оглавление

Машинный перевод

История

Перевод на основе правил

- Системы дословного перевода

- Трансферные системы

- Интерлингвистические системы

- Преимущества и недостатки RBMT

Перевод на основе примеров

- Близость предложений

- Рекомбинация

- Преимущества и недостатки EBMT

Статистический перевод

- Статистический перевод по словам

- Статистический перевод по фразам

- SMT на основе синтаксиса

- Преимущества и недостатки SMT

Донейросетевой МП: что сейчас и что дальше?

Определение

Машинный перевод - процесс перевода с одного естественного языка на другой с помощью специальной компьютерной программы.

В чем польза?

- ▶ Повышение эффективности труда переводчиков
- ▶ Единство терминологии и стиля - уменьшение затрат на редакторскую правку
- ▶ Человек не обязан знать язык перевода

История

- ▶ 1933г, СССР: машина Петра Троянского

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

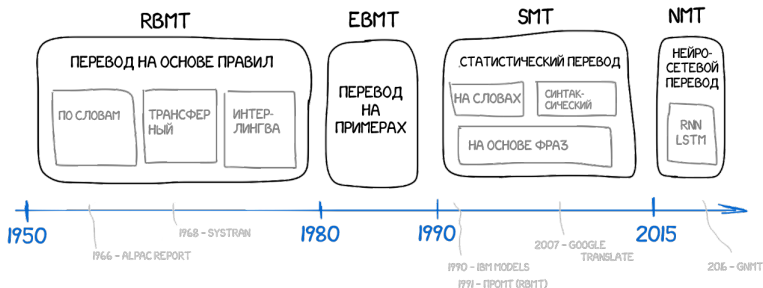


- ▶ 7 января 1954г, штаб-квартира IBM, Нью-Йорк:
Джорджтаунский эксперимент, компьютер IBM 701



Основные подходы

- ▶ МП на основе правил (Rule-Based, RBMT)
- ▶ МП на примерах (Example-Based, EBMT)
- ▶ Статистический (Statistical, SMT)
- ▶ Нейросетевой (Neural, NMT)



Оглавление

Машинный перевод

История

Перевод на основе правил

Системы дословного перевода

Трансферные системы

Интерлингвистические системы

Преимущества и недостатки RBMT

Перевод на основе примеров

Близость предложений

Рекомбинация

Преимущества и недостатки EBMT

Статистический перевод

Статистический перевод по словам

Статистический перевод по фразам

SMT на основе синтаксиса

Преимущества и недостатки SMT

Донейросетевой МП: что сейчас и что дальше?

Основная идея RBMT

Имитация действий переводчика:

- ▶ **двуязычный словарь** (напр, EN -> RU)
- ▶ **набор лингвистических правил под каждый язык** (напр, RU: существительные женского рода оканчиваются на -а/-я)
- ▶ **дополнительные списки имён, корректоры орфографии, транслитераторы и т.д.**



Виды RBMT

- ▶ Системы пословного перевода (Direct MT)
- ▶ Трансферные системы (Transfer-based MT)
- ▶ Интерлингвистические системы (Interlingual MT)

Системы дословного перевода

- ▶ Пословный перевод + правка морфологии, согласование падежей, окончания и остальной синтаксис
- ▶ Лингвистами прописываются правила под каждое слово
- ▶ В результате очень низкое качество
- ▶ В современных системах подход не используется

я	ХОЧУ	СОРОК	КИЛОГРАММ	ХУРМЫ
↓	↓	↓	↓	↓
I	WANT	FORTY	KILOGRAM	PERSIMMONS

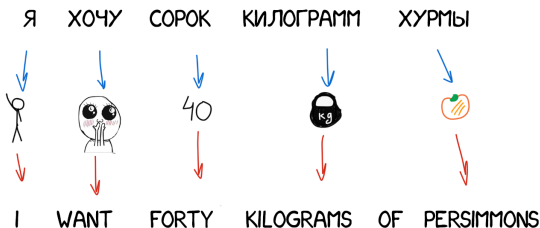
Трансферные системы

- ▶ Три этапа: анализ, трансфер и синтез. *Пример: PROMT*
- ▶ Выделение синтаксических конструкций (подлежащее, сказуемое и т.д.). Не закладываются правила перевода каждого слова, манипулирование целыми конструкциями
- ▶ **Преимущества:** можно задать общие правила согласования по роду и падежу + в теории можно добиться хорошей конвертации порядка слов в языках
- ▶ **Недостатки:** комбинаций слов больше, чем самих слов + все еще много работы лингвистов + все еще низкое качество



Интерлингвистические системы

- ▶ Два этапа: **анализ** и **синтез** на основе правил и словарей соответствующих языков
- ▶ Для перевода с одного языка на другой используется промежуточное представление interlingua (метаязык с едиными правилами для всех существующих языков)
- ▶ **Преимущества:** в отличие от трансферной системы, можем свободно добавлять языки к уже имеющимся
- ▶ **Недостатки:** создать универсальную интерлингву вручную оказалось крайне сложно



Преимущества и недостатки RBMT

► Преимущества

- Синтаксическая и морфологическая точность (не путает слова)
- Стабильность и предсказуемость результата (все переводчики получают одинаковый результат)
- Возможность настройки на предметную область

► Недостатки

- Трудоемкость и длительность разработки (учет всех исключений из правил + омонимия)
- Необходимость поддерживать и актуализировать лингвистические базы данных
- «Машинный акцент» при переводе:
Since the Desert One debacle, the United States has poured vast resources into its special forces.
Начиная с разгрома Пустыни Один, Соединенные Штаты вылили обширные ресурсы в свой спецназ

Оглавление

- Машинный перевод

 - История

- Перевод на основе правил

 - Системы дословного перевода

 - Трансферные системы

 - Интерлингвистические системы

 - Преимущества и недостатки RBMT

- Перевод на основе примеров

 - Близость предложений

 - Рекомбинация

 - Преимущества и недостатки EBMT

- Статистический перевод

 - Статистический перевод по словам

 - Статистический перевод по фразам

 - SMT на основе синтаксиса

 - Преимущества и недостатки SMT

- Донейросетевой МП: что сейчас и что дальше?

Основная идея ЕВМТ

- ▶ В 1984 году учёному университета Киото по имени Макото Нагао приходит идея: а что если не пытаться каждый раз переводить заново, а использовать уже готовые фразы?
- ▶ Четыре этапа: мэтчинг фраз с фразами из базы темплейтов с помощью метрик близости, перевод фрагментов, рекомбинация, выравнивание (постобработка - например, согласование существительного с глаголом)
- ▶ Для подбора близких фрагментов необходимы: метрика близости и достаточно объемная база темплейтов (WordNet с указанием гиперонимов, "sisters" или Wikipedia)

Я ИДУ В ТЕАТР = I'M GOING TO THE THEATER

Я ИДУ В МАГАЗИН ^{???} = I'M GOING TO THE STORE

STORE

Близость предложений

Word-based similarity

- ▶ Расстояние редактирования (расстояние Левенштейна)
- ▶ Для векторного представления предложений мешок слов + индекс Дайса или мера Жаккара для вычисления близости векторов
- ▶ tf-idf + косинусный коэффициент

Пример: косинусный коэффициент + tf-idf

S_1 = "Peter hired a car for the trip."

S_2 = "For the trip, a car was hired by Peter."

$V(S_1) = tf(S_1) \cdot idf(S_1)$ - векторное представление S_1

$V(S_2) = tf(S_2) \cdot idf(S_2)$ - векторное представление S_2

$\cosine(V(S_1), V(S_2)) = \frac{V(S_1) \cdot V(S_2)}{|V(S_1)| \cdot |V(S_2)|}$ - мера близости S_1 и S_2

Близость предложений

Tree and graph-based similarity

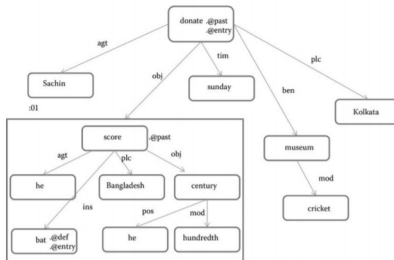
► Близость деревьев разбора:

Based on constituency and dependency parse trees of S_1 and S_2 .

(ROOT
(S
(NP (PRP\$ My) (NN dog))
(ADVP (RB also))
(VP (VBZ likes)
(S
(VP (VBG eating)
(NP (NN sausage))))
(.)))

nmod:poss(dog-2, My-1)
nsubj(likes-4, dog-2)
advmod(likes-4, also-3)
root(ROOT-0, likes-4)
xcomp(likes-4, eating-5)
dobj(eating-5, sausage-6)

► Близость семантических графов:



Близость предложений

Tree and graph-based similarity

Пример: constituency tree similarity

N_1 - количество вершин в S_1

N_2 - количество вершин в S_2

M - количество смэтчившихся вершин при заданном порядке обхода деревьев

$$S(S_1, S_2) = \frac{M}{\max(N_1, N_2)}$$

Если $S >$ определенного порога, деревья считаются близкими

```
(ROOT
  (S
    (NP (PRP He))
    (VP (VBZ buys)
      (NP
        (NP (DT a) (NN book))
        (PP (IN on)
          (NP (JJ international) (NNS politics))))))
    (. )))
```

```
(ROOT
  (S
    (NP (PRP He))
    (VP (VBZ buys)
      (NP (NNS mangoes))))
  (. )))
```

Рекомбинация

Фрагменты текста, извлечённые на этапе соответствий, объединяются для создания целого предложения.

Способы:

- ▶ Based on sentence parts
- ▶ Based on properties of sentence parts - properties can be features such as word, lemma, gender, number (singular/plural), person (3rd, 2nd), tense (past,future), voice (passive,active), POS tag, etc.
- ▶ Based on parts of semantic graphs

Рекомбинация

Пример: основанная на частях предложения

Input: "Tomorrow, today will be yesterday."

Example: "Yesterday, today was tomorrow."

(Tomorrow, today and yesterday are hyponyms of day.)

(will be and was both derived from the verb to be)

Input: "Tomorrow, today will be yesterday."



Example: "Yesterday, today was tomorrow."

Example translation: "Gerstern, heute war morgen."



Output translation: "Morgen, heute ist gerstern."

Преимущества и недостатки ЕВМТ

- ▶ Преимущества:

- ▶ Движение в сторону работы с многозначностью слов:
например, хорошо справляется с фразовыми глаголами в английском языке

1) Ram **put on** the lights. (Switched on) (перевод на хинди—урду: *Jalana*)

2) Ram **put on** a cap. (Wear) (перевод на хинди—урду: *Rahenna*)

- ▶ Недостатки:

- ▶ Необходима подробная и объемная база темплейтов

Оглавление

- Машинный перевод

 - История

- Перевод на основе правил

 - Системы дословного перевода

 - Трансферные системы

 - Интерлингвистические системы

 - Преимущества и недостатки RBMT

- Перевод на основе примеров

 - Близость предложений

 - Рекомбинация

 - Преимущества и недостатки EBMT

- Статистический перевод

 - Статистический перевод по словам

 - Статистический перевод по фразам

 - SMT на основе синтаксиса

 - Преимущества и недостатки SMT

- Донейросетевой МП: что сейчас и что дальше?

Основная идея SMT

- ▶ На рубеже 1990 года в исследовательском центре IBM впервые показали систему МП, которая ничего не знала о правилах и лингвистике
- ▶ Среди многочисленных переводов (параллельные корпуса) система находит наиболее популярный, его и предоставляя в качестве ответа
- ▶ $f = f_1, f_2, \dots, f_n$ - французское предложение
 $e = e_1, e_2, \dots, e_m$ - его английский перевод
Хотим: $\max_e P(e|f) = \frac{P(f|e) \cdot P(e)}{P(f)}$ по всем возможным e
То есть ищем: $e' = \operatorname{argmax}_e (P(f|e) \cdot P(e))$
- ▶ $P(f|e)$ - Translation model, $P(e)$ - Language model
- ▶ Компоненты: Translation model, Language model, Decoder



Language model

- ▶ Показывает, насколько корректно предложение в рамках своего языка

- ▶ $P(e) = \prod_{i=1}^m P(e_i | e_1, \dots, e_{i-1})$ - при больших i почти 0.
"unseen" \neq "impossible"

- ▶ Приближение n-gram language model:

$$P(e) = \prod_{i=n}^m P(e_i | e_{i-n+1}, \dots, e_{i-1})$$

- ▶ $P(e_i | e_{i-n+1}^{i-1}) = \frac{\#(e_{i-n+1}, \dots, e_i)}{\#(e_{i-n+1}, \dots, e_{i-1})}$, где $\#$ - количество фраз
- ▶ Но для некоторых фраз и при малых n P почти 0.

Smoothing:

$$P_{smooth}(e_i | e_{i-n+1}^{i-1}) = P^*(e_i | e_{i-n+1}^{i-1}), \text{ if } \#(e_{i-n+1}^{i-1}) > 0 \\ = \alpha(e_i | e_{i-n+1}^{i-1}) \cdot P_{smooth}(e_i | e_{i-n+2}^{i-1}), \text{ otherwise}$$

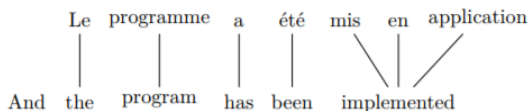
where P^* is computed by discounting,

$$\alpha(e_i | e_{i-n+1}^{i-1}) = \frac{1 - \sum_{e: \# > 0} P(e_i | e_{i-n+1}^{i-1})}{1 - \sum_{e: \# > 0} P(e_i | e_{i-n+2}^{i-1})}$$

Translation model

- ▶ $P(f|e)$ - вероятность того, что e и f - действительно пара по переводу
- ▶ Введем "alignment": $a = (a_1, a_2, \dots, a_s)$, где a_i - номер слова в исходном предложении, которому при переводе соответствует i -е слово в переведенном предложении

Пример:



$$a_1, a_2, \dots, a_7 = \langle 2, 3, 4, 5, 6, 6, 6 \rangle$$

- ▶ Тогда вместо моделирования $P(f|e)$ мы можем моделировать alignment model: $P(e|f) = \sum_a P(f, a|e)$

Виды SMT

- ▶ Word-based SMT
- ▶ Phrase-based SMT
- ▶ Syntax-based SMT

Статистический перевод по словам

- ▶ Первая модель: **IBM model 1**
- ▶ Пословный подбор наиболее вероятного перевода
- ▶ Нет учета порядка слов в переводе
- ▶ Зато модель умела переводить слово в конструкцию из нескольких (но не факт, что обратно): Der Staubsauger (пылесос) -> Vacuum Cleaner

Я БОЛЬШЕ НЕ ХОЧУ ХУРМЫ

| 0.88 / 0.45 / 0.79 / 0.81 | 0.91

I MORE NOT WANT PERSIMMONS

Статистический перевод по словам

- ▶ IBM model 2
- ▶ Добавился промежуточный шаг: после перевода машина пыталась переставить слова местами так, как она думала будет звучать более естественно

Я БОЛЬШЕ НЕ ХОЧУ ХУРМЫ

| 0.88 / 0.45 / 0.79 / 0.81 | 0.91

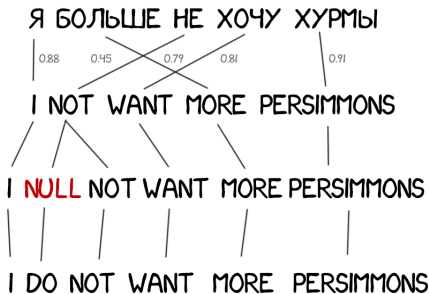
I MORE NOT WANT PERSIMMONS

| |

I NOT WANT MORE PERSIMMONS

Статистический перевод по словам

- ▶ IBM model 3
- ▶ Часто при переводе появляются новые слова, которых не было в оригинальном тексте
- ▶ Добавилось два промежуточных шага:
 - ▶ Вставка маркеров (NULL-слов) на те места, где машина подозревает необходимость нового слова
 - ▶ Подбор нужного артикля, частицы или глагола под каждый маркер



Статистический перевод по словам

► ISM model 4:

- Учет "относительного порядка": запоминаются слова, которые при переводе меняются местами (например, некоторые прилагательные + существительные при переводе с английского на французский)

► IBM model 5:

- Добавили параметров для обучения
- Пофиксили проблемы, когда два слова конфликтовали за место в предложении

Модели статистического перевода по словам не могли справиться с омонимией, падежами и родом - не учитывали контекст.

Статистический перевод по фразам

- ▶ Для обучения текст разбивался не только на слова, но и на N-граммы
- ▶ Увеличение точности перевода
- ▶ Нестабильность перевода и случаи перевода вида «three hundred» -> «300»
- ▶ До 2016-го года Google Translate, Yandex, Bing и другие качественные онлайн-переводчики работали именно как Phrase-based

FULL SUPERIORITY OF PERSIMMONS

ПЕРЕВОД ПО СЛОВАМ
(ХОРОШО, НО ДОСЛОВНО)

ПОЛНОЕ ПРЕВОСХОДСТВО ХУРМЫ

COMPLETE SUPERIORITY

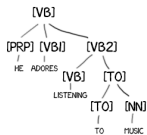
PERSIMMON SUPERIORITY

ПЕРЕВОД ПО ФРАЗАМ
(УЧИТЫВАЕТ КОНТЕКСТ
СОСЕДНИХ СЛОВ)

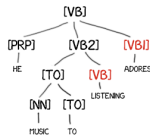
COMPLETE PERSIMMON SUPERIORITY

SMT на основе синтаксиса

HE ADORES LISTENING TO MUSIC

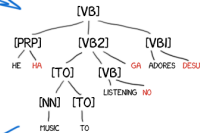


ВХОДНОЕ ПРЕДЛОЖЕНИЕ



КОНВЕРТАЦИЯ
ДЕРЕВА

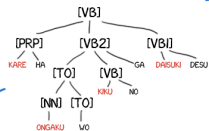
СИНТАКСИЧЕСКИЙ
ПЕРЕВОД



ВСТАВКА СЛОВ

KARE HA ONGAKU WO KIKU NO GA DAISUKI DESU

РЕЗУЛЬТАТ



ПЕРЕВОД

Преимущества и недостатки SMT

- ▶ Преимущества:
 - ▶ Быстрая настройка
 - ▶ Легкость добавления новых направлений перевода
 - ▶ Гладкость перевода
 - ▶ Не требуется работа лингвистов
- ▶ Недостатки:
 - ▶ Дефицит параллельных корпусов
 - ▶ Многочисленные грамматические ошибки
 - ▶ Большая нестабильность перевода

Оглавление

- Машинный перевод

 - История

- Перевод на основе правил

 - Системы дословного перевода

 - Трансферные системы

 - Интерлингвистические системы

 - Преимущества и недостатки RBMT

- Перевод на основе примеров

 - Близость предложений

 - Рекомбинация

 - Преимущества и недостатки EBMT

- Статистический перевод

 - Статистический перевод по словам

 - Статистический перевод по фразам

 - SMT на основе синтаксиса

 - Преимущества и недостатки SMT

- Донейросетевой МП: что сейчас и что дальше?






Донейросетевой МП: Что сейчас и что дальше?

- ▶ Донейросетевые методы чаще всего комбинируются в Гибридные системы, нивелируя недостатки друг друга
- ▶ Использование нейросетей для решения задачи машинного перевода
- ▶ Машинный перевод все еще уступает по качеству человеческому переводу
- ▶ Ограниченное количество параллельных корпусов для SMT и NMT

Выводы

- ▶ RBMT имитирует действия лингвиста, использует словарь и набор правил
- ▶ EBMT не совершает глубокого лингвистического анализа, но собирает перевод предложения из примеров из базы темплейтов
- ▶ SMT подбирают наиболее популярный перевод по параллельным корпусам текстов
- ▶ В чистом виде донейросетевые методы почти не встретишь - соединяются в Гибридные системы МП

Источники I

-  https://vas3k.ru/blog/machine_translation/
-  http://www.promt.ru/images/ainl_molchanov_promt.pdf
-  Подробнее о мерах близости и рекомбинации предложений при EBMT (слайды 35-63):
https://hpi.de/fileadmin/user_upload/fachgebiete/plattner/teaching/MachineTranslation/MT2016/MT13_ExampleBasedMT.pdf
-  Language and Translation models:
<http://michaelnielsen.org/blog/introduction-to-statistical-machine-translation/>
-  Translation model:
http://www.cs.sfu.ca/~anoop/students/anahita_mansouri/anahita-depth-report.pdf

Источники II



Smoothing:

<https://cxwangyi.wordpress.com/2010/07/28/backoff-in-n-gram-language-models/>



http://www.machinelearning.ru/wiki/images/5/5d/Mel_lain_msu_nlp_sem_2.pdf



<https://homepages.inf.ed.ac.uk/pkoehn/publications/tutorial2006.pdf>



https://ru.wikipedia.org/wiki/Машинный_перевод_на_основе_примеров



<https://moluch.ru/conf/phil/archive/138/8497/>