

Практический анализ данных и машинное обучение: искусственные нейронные сети

Ульянкин Филипп, Соловей Влад

30 мая 2019 г.

Seq2Seq модели

Agenda

- История автоперевода
- Нейросетевой перевод
- Генерация подписи по картинке

История автоперевода

✖️ 🔍 🎧 📄 A• АНГЛИЙСКИЙ↔️РУССКИЙ

Введите текст или адрес сайта

0 / 10000

✖️ 🔍 📄 ↗️ ⌂ ⌂ ⌂ ⌂ ⌂ ⌂

Google Переводчик

≡

Text Documents

ОПРЕДЕЛИТЬ ЯЗЫК АНГЛИЙСКИЙ РУССКИЙ НЕМЕЦКИЙ ↔️ РУССКИЙ АНГЛИЙСКИЙ УКРАИНСКИЙ

Перевод

🔊 0/5000

Отправить отзыв

Когда и кто захотел

- Холодная война, 1954 год. США и СССР хотят получить машину для автопереводов
- Джорджтаунский эксперимент. Перевод компом IBM 60 предложений с перфокарт ⇒ приток денег в исследования
- В СССР аналогичный эксперимент
- В обоих государствах все предложения для перевода тщательно подобраны и оттестированы :(

Когда и кто захотел

- Учёные обещают, что в течение ближайших 5 лет задача машинного перевода будет решена.
- Через 12 лет, в 1966 американский комитет ALPAC публикует отчёт, в котором называет машинный перевод дорогим, неточным и бесперспективным

Перевод на основе правил (Rule-based machine translation, RBMT)

- Получает интенсивное развитие в 1970-х годах
- Словари + попытка посмотреть на то как работают лингвисты и вбить какие-то паттерны в компьютер (существительные оканчиваются на а-я и тп)
- Бывает разных видов

- Дословный перевод (Direct Machine Translation): делим текст по словам, переводим каждое, правим каждое слово в соответствии с накопленными правилами (окончания, падежи и тп). Правила придумывают лингвисты.
- Трансферные системы (Transfer-based Machine Translation): сначала выделяем синтаксические конструкции (сказуемые, подлежащие и тп), понимаем как слова надо переставить, а потом уже переводим.

Эти типы стали есть на складе

- Придумывать правила вручную - сложно
- Огромное количество исключений
- Омонимия (разный смысл одних и тех же слов в зависимости от контекста)
- RBMT системы за годы холодной войны вышли на пик и успешно умерли, сегодня они не используются нигде

Example based machine translation (EBMT)

- Япония, 1984 г.
- Зачем каждый раз переводить заново? Давайте переиспользовать!
- Впервые возникла идея просто скармливать компьютеру существующие данные, а не придумывать правила

Example of bilingual corpus

English

How much is that **red umbrella**? Ano **akai kasa** wa ikura desu ka.

Japanese

How much is that **small camera**? Ano **chiisai kamera** wa ikura desu ka.

Statistical machine translation (SMT)

- IBM, 1990 г.
- Берём корпус параллельных текстов, смотрим как часто слово house переводится как дом, строение, постройка, так и переводим!
- Работало лучше, чем всё что было до этого

ОПРЕДЕЛИТЬ ЯЗЫК
АНГЛИЙСКИЙ
РУССКИЙ
НЕМЕЦКИЙ
▼
+
-
РУССКИЙ
АНГЛИЙСКИЙ
УКРАИНСКИЙ
▼

house

X

ЖИЛОЙ ДОМ

zhiloy dom

☆

5/5000
...

house – определения

Имя Существительное

① a building for human habitation, especially one that is lived in by a family or small group of people.
 "Real foxes do, indeed, sometimes make their homes under human houses and, increasingly in this country at any rate, under city homes."

Синонимы:

residence home place of residence homestead a roof over one's head
 habitation dwelling (place) abode domicile

Глагол

① provide (a person or animal) with shelter or living quarters.
 "attempts by the government to house the poor"

Синонимы:

accommodate provide accommodations for
 give someone a roof over their head lodge quarter board billet take in

house: варианты перевода

Имя Существительное Частота ⓘ

дом	house, home, dwelling, door, premises, crib	
жилище	housing, home, dwelling, house, abode, habitation	
театр	theater, house, playhouse, stage, living theater, theatre	
здание	building, house, structure, edifice, construction, fabric	
палата	chamber, ward, house	
гостиница	hotel, inn, hostel, guesthouse, pub, house	
семья	family, household, home, kin, colony, house	
рубка	felling, cutting, deckhouse, house, chopping	
хозяйство	economy, farm, household, property, house, establishment	
род	genus, race, kind, family, type, house	

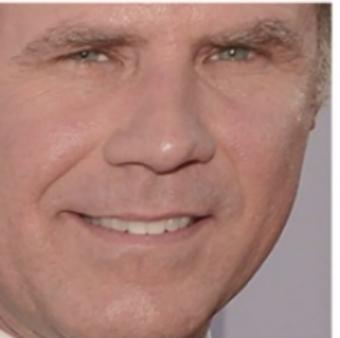
Word-based SMT

- Начали конечно же со статистического перевода по отдельным словам
- Пример реализации на python: <https://github.com/shawa/IBM-Model-1>
- Дальше попробовали также по статистике переставлять слова, добавлять недостающие артикли
- Всё ещё много проблем с омонимией и согласованностью слов в предложениях

Phrase-based SMT

- Word-based SMT был основан на мешке слов, тут подмешали N-граммы
- С 2006 года этот подход использовали абсолютно все
- Так продолжалось до 2016 года
- В 2016 году Google перевернул игру:
<https://ai.googleblog.com/2016/09/a-neural-network-for-machine.html>

Input Image



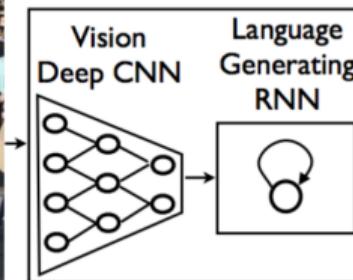
128 Measurements Generated from Image

0.097496084868908	0.045223236083984	-0.1281466782093	0.032084941864014
0.12529824674129	0.060309179127216	0.17521631717682	0.020976085215807
0.030809439718723	-0.01981477253139	0.10801389068365	-0.00052163278451189
0.036050599068403	0.065554238855839	0.0731306001544	-0.1318951100111
-0.097486883401871	0.122662897253	-0.029626874253154	-0.0059557510539889
-0.0066401711665094	0.036750309169292	-0.15958009660244	0.043374512344599
-0.14131525158862	0.14114324748516	-0.031351584941149	-0.053343612700701
-0.048540540039539	-0.061901587992907	-0.15042643249035	0.078198105096817
-0.12567175924778	-0.10568545013666	-0.12728653848171	-0.076289616525173
-0.061418771743774	-0.074287034571171	-0.065365232527256	0.12369467318058
0.046741496771574	0.0061761881224811	0.14746543765068	0.056418422609568
-0.12113650143147	-0.21055991947651	0.0041091227903962	0.089727647602558
0.061606746166945	0.11345765739679	0.021352224051952	-0.0085843298584223
0.061989940702915	0.19372203946114	-0.086726233363152	-0.022388197481632
0.10904195904732	0.084853030741215	0.09463594853878	0.020696049556136
-0.019414527341723	0.0064811296761036	0.21180312335491	-0.050584398210049
0.15245945751667	-0.16582328081131	-0.035577941685915	-0.072376452386379
-0.12216668576002	-0.0072777755558491	-0.036901291459799	-0.034365277737379
0.083934650121613	-0.059730989369411	-0.070026844739914	-0.045013956725597
0.087945111095905	0.11478432267904	-0.089621491730213	-0.013955107890069
-0.021407851949334	0.14841195940971	0.078333757817745	-0.17898085713387
-0.018298890441656	0.049525424838066	0.13227833807468	-0.072600327432156
-0.011014151386917	-0.051016297191381	-0.14132921397686	0.0050511928275228
0.0093679334968328	-0.062812767922878	-0.13407498598099	-0.014829395338893
0.058139257133007	0.0048638740554452	-0.039491076022387	-0.043765489012003
-0.024210374802351	-0.11443792283535	0.071997955441475	-0.012062266469002
-0.057223934680223	0.014683869667351	0.05228154733777	0.012774495407939
0.023535015061498	-0.081752359867096	-0.031709920614958	0.069833360612392
-0.0098039731383324	0.037022035568953	0.11009479314089	0.11638788878918
0.020220354199409	0.12788131833076	0.18632389605045	-0.015336792916059
0.0040337680839002	-0.094398014247417	-0.11768248677254	0.10281457751989
0.051597066223621	-0.10034311562777	-0.040977258235216	-0.082041338086128

Measurements Generated from Sentence			
Input Sentence			
"Machine Learning is Fun!" →			
0.097496084868908	0.0452232360833984	-0.1281466782093	0.032084941864014
0.12529824674129	0.060309179127216	0.17521631717682	0.020976085215807
0.030809439718723	-0.01981477253139	0.10801389068365	-0.00052163278451189
0.03605059068403	0.065554238855839	0.0731306001544	-0.1318951100111
-0.097486883401871	0.1226262897253	-0.029626874253154	-0.0059557510539889
-0.0066401711665094	0.036750309169292	-0.15958009660244	0.043374512344599
-0.14131525158882	0.14114324748516	-0.031351584941149	-0.053343612700701
-0.048540540039539	-0.061901587992907	-0.15042643249035	0.078198105066817
-0.125671715924778	-0.10568545013666	-0.12728653848171	-0.076289616525173
-0.061418771743774	-0.074287034571171	-0.065365232527256	0.12369467318058
0.046741496771574	0.0061761881242811	0.14746543765068	0.056418422609568
-0.12113650143147	-0.21055991947651	0.0041091227903962	0.089727647602558
0.061606746166945	0.11345765739679	0.021352224051952	-0.00858432985854223
0.061989940702915	0.19372203946114	-0.086726233363152	-0.022388197481632
0.10904195904732	0.084853030741215	0.09463594853878	0.020696049556136
-0.019414527341723	0.0064811296761036	0.21180312335491	-0.050584398210049
0.15245945751667	-0.16582328081131	-0.035577941685915	-0.072376452386379
-0.12216668576002	-0.007277755558491	-0.036901291459799	-0.034365277737379
0.083934605121613	-0.059730969369411	-0.070026844739914	-0.045013956725597
0.087945111095905	0.11478432267904	-0.089621491730213	-0.013955107890069
-0.021407851949334	0.14841195940971	0.078333757817745	-0.17898085713387
-0.018298890441656	0.049525424838066	0.13227833807468	-0.072600327432156
-0.011014151386917	-0.051016297191381	-0.14132921397686	0.0050511928275228
0.0093679334968328	-0.062812767922878	-0.13407498598099	-0.014829395338893
0.0581392571133007	0.0048638740554452	-0.039491076022387	-0.043785489012003
-0.024210374802351	-0.11443792283535	0.071997955441475	-0.012062266469002
-0.057223934680223	0.014683869667351	0.05228154733777	0.012774495407939
0.023535015061498	-0.081752359867096	-0.031709920614958	0.069833360612392
-0.0098039731383324	0.037022035568953	0.11009479314089	0.11638788878918
0.0202220354199409	0.12788131833076	0.18632389605045	-0.015336792916059
0.0040337680839002	-0.094398014247417	-0.11768248677254	0.10281457751989
0.051597066223621	-0.10034311562777	-0.040977258235216	-0.082041338086128

Текст + картинка

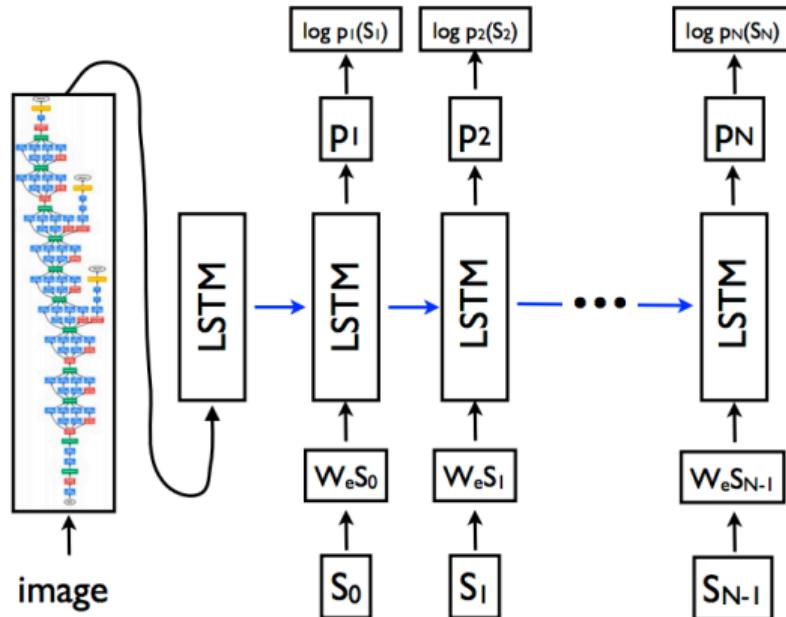
Генерация подписи по картинке



**A group of people
shopping at an
outdoor market.**

**There are many
vegetables at the
fruit stand.**

Генерация подписи по картинке



<https://arxiv.org/abs/1411.4555>

Генерация подписи по картинке

A person riding a motorcycle on a dirt road.



Two dogs play in the grass.



A skateboarder does a trick on a ramp.



A dog is jumping to catch a frisbee.



A group of young people playing a game of frisbee.



Two hockey players are fighting over the puck.



A little girl in a pink hat is blowing bubbles.



A refrigerator filled with lots of food and drinks.



A herd of elephants walking across a dry grass field.



A close up of a cat laying on a couch.



A red motorcycle parked on the side of the road.



A yellow school bus parked in a parking lot.



Describes without errors

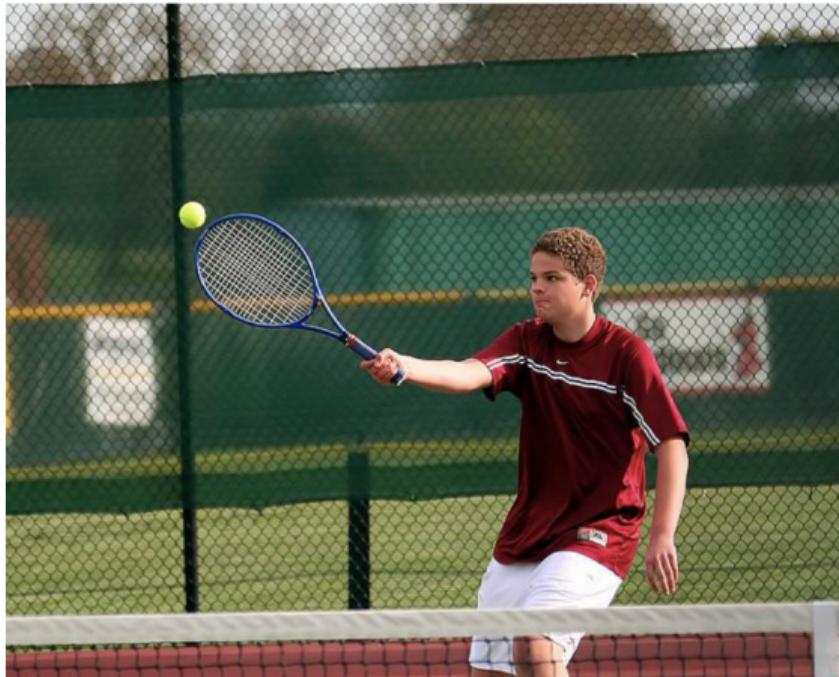
Describes with minor errors

Somewhat related to the image

Unrelated to the image

Figure 5. A selection of evaluation results, grouped by human rating.

Генерация подписи по картинке



a man is playing tennis on a tennis court

Генерация подписи по картинке



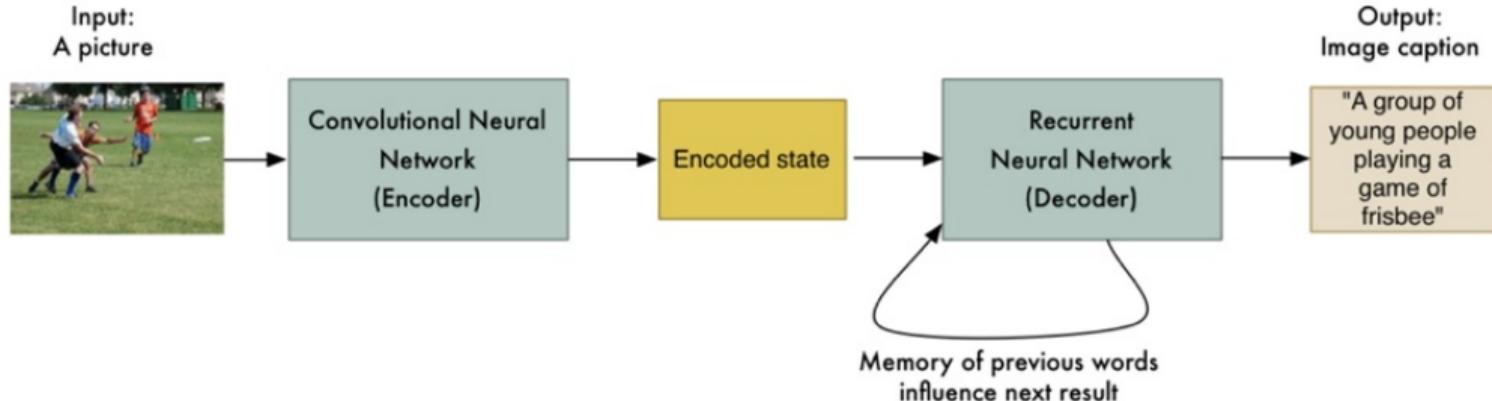
a man sitting on a bus with a dog

Генерация подписи по картинке



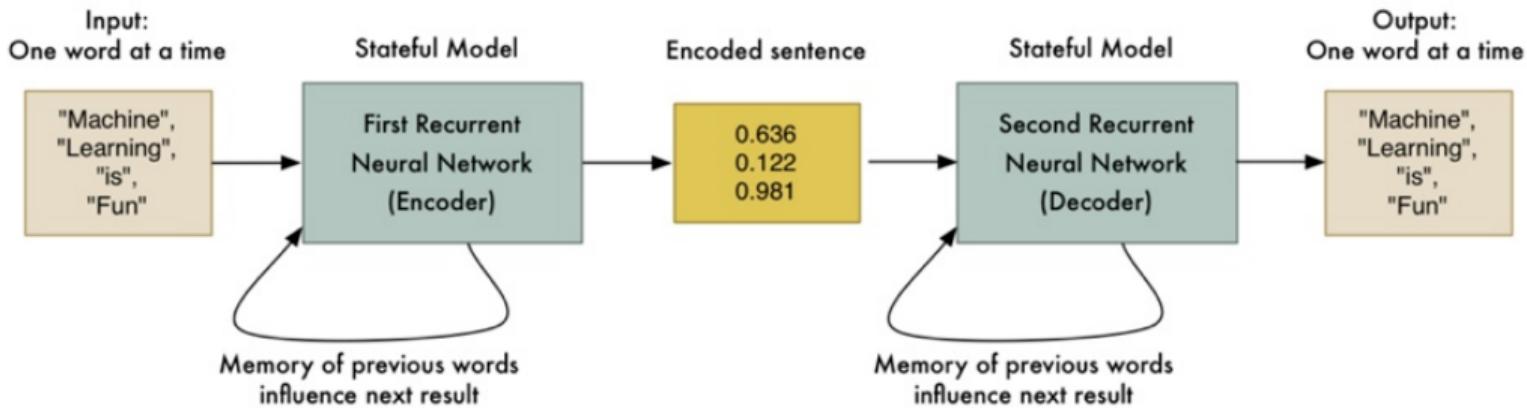
a bunch of bananas are hanging from a ceiling

Архитектура

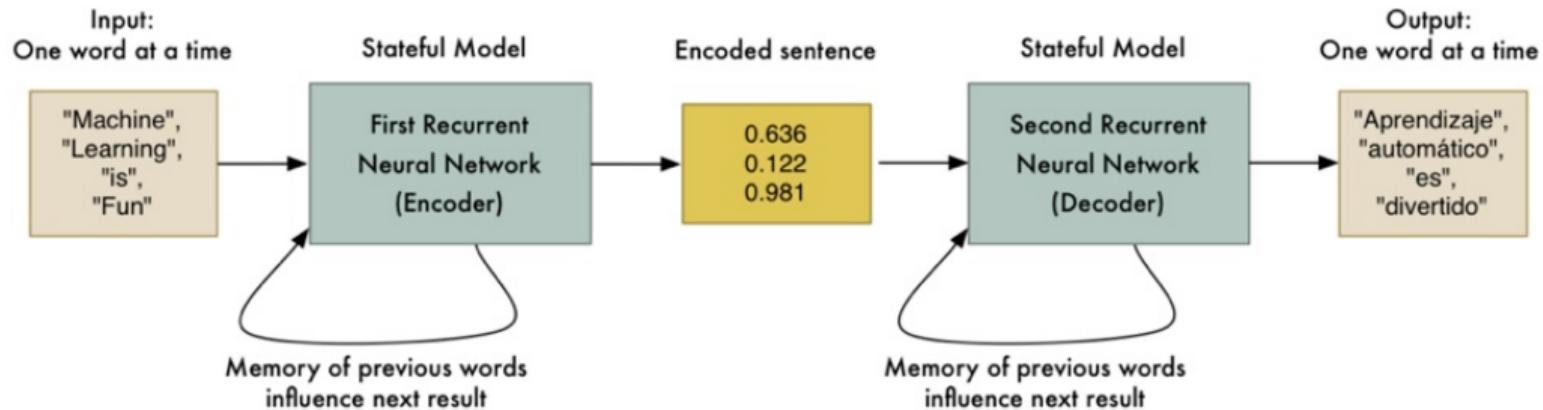


Автопереводчик

Автокодировщик для текстов



Переводчик



Переводчик

- Архитектуры переводчиков очень разные
- Поначалу это были RNN, после перешли на многослойные двунаправленные LSTM
- Сегодня есть ещё куча разных штук, которыми пичкают автопереводчики
- Например, пробуют разные виды эмбедингов, добавляют модели с вниманием для согласования контекста

Google vs Yandex

- Переводчик Google - нейросетевой
- Переводчик Yandex очень сильно от него отличается
- На коротких фразах нейросетевой перевод работает плохо, поэтому
- Перевод идёт двумя способами: статистическим и нейросетевым, а после Catboost выбирает тот, который лучше подходит

Ссылки

- Пресс-релиз Яндекса от 2017: <https://yandex.ru/blog/company/kak-pobedit-mornikov-yandeks-zapustil-gibridnuyu-sistemu-perevoda>
- Статья Яндекса на хабре про историю автоперевода:
<https://habr.com/ru/company/yandex/blog/224445/>

Нейросети - Lego

