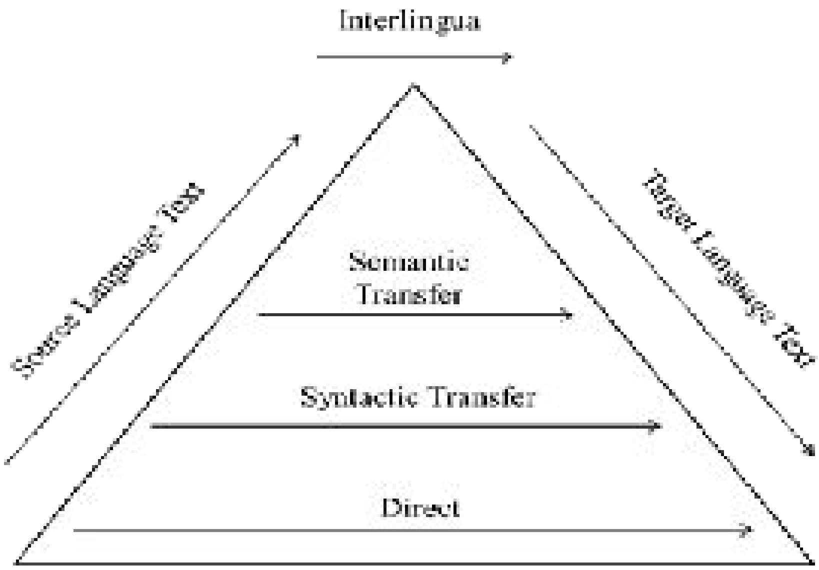
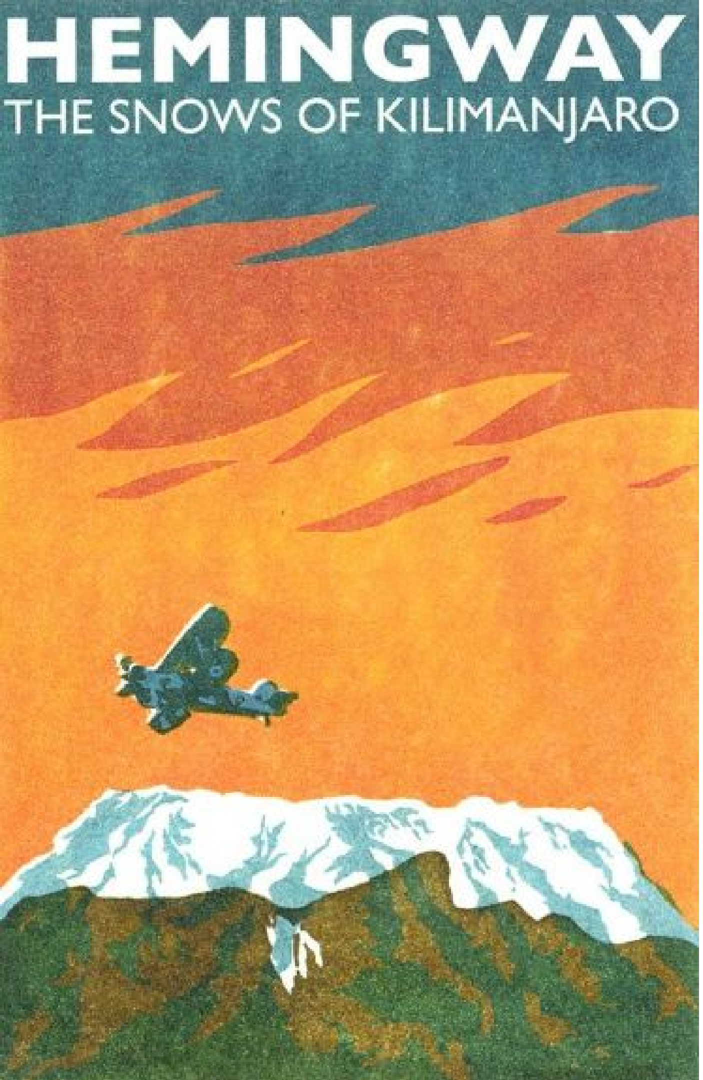
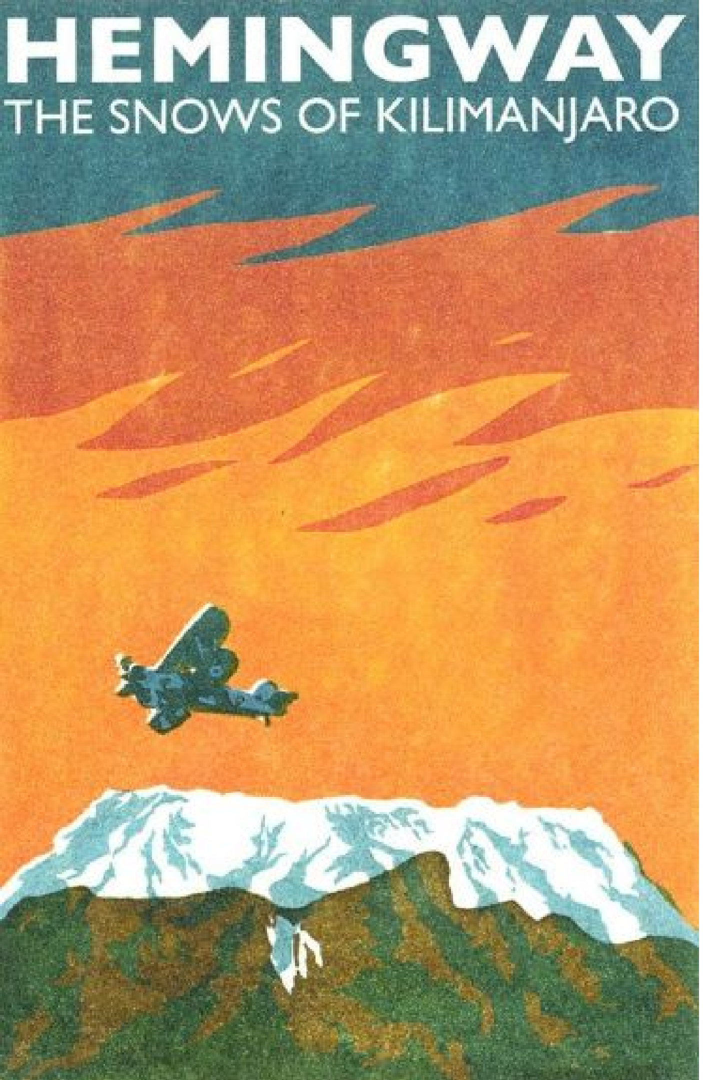
70 Years of Machine Translation





Original

Kilimanjaro is a snow-covered mountain 19,710 feet high, and is said to be the highest mountain in Africa. Its western summit is called the Masai "Ngaje Ngai," the House of God. Close to the western summit there is the dried and frozen carcass of a leopard. No one has explained what the leopard was seeking at that altitude.

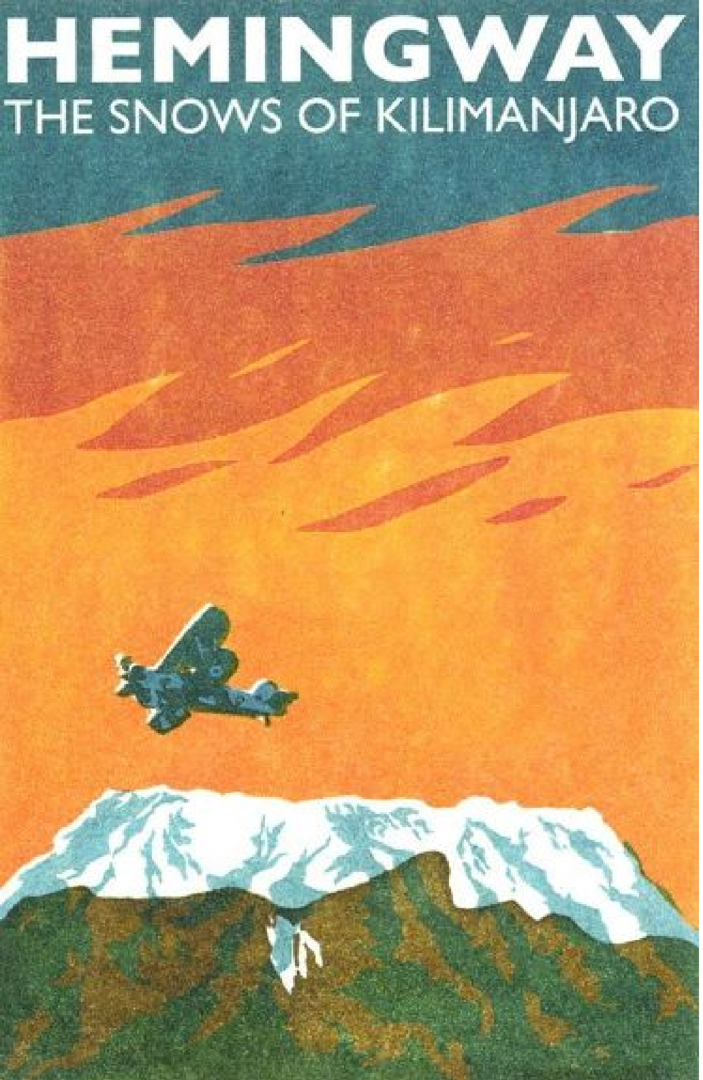


Original

Kilimanjaro is a snow-covered mountain 19,710 feet high, and is said to be the highest mountain in Africa. Its western summit is called the Masai "Ngaje Ngai," the House of God. Close to the western summit there is the dried and frozen carcass of a leopard. No one has explained what the leopard was seeking at that altitude.

Back translation from Japanese (old)

Kilimanjaro is 19,710 feet of the mountain covered with snow, and it is said that the highest mountain in Africa. Top of the west, "Ngaje Ngai" in the Maasai language, has been referred to as the house of God. The top close to the west, there is a dry, frozen carcass of a leopard. Whether the leopard had what the demand at that altitude, there is no that nobody explained.



Original

Kilimanjaro is a snow-covered mountain 19,710 feet high, and is said to be the highest mountain in Africa. Its western summit is called the Masai "Ngaje Ngai," the House of God. Close to the western summit there is the dried and frozen carcass of a leopard. No one has explained what the leopard was seeking at that altitude.

Back translation from Japanese (old)

Kilimanjaro is 19,710 feet of the mountain covered with snow, and it is said that the highest mountain in Africa. Top of the west, "Ngaje Ngai" in the Maasai language, has been referred to as the house of God. The top close to the west, there is a dry, frozen carcass of a leopard. Whether the leopard had what the demand at that altitude, there is no that nobody explained.

Back translation from Japanese (new)

Kilimanjaro is a mountain of 19,710 feet covered with snow, which is said to be the highest mountain in Africa. The summit of the west is called "Ngaje Ngai" God's house in Masai language. There is ~~a~~ dried and frozen carcass of a leopard near the summit of the west. No one can explain what ~~the~~ leopard was seeking at that altitude.

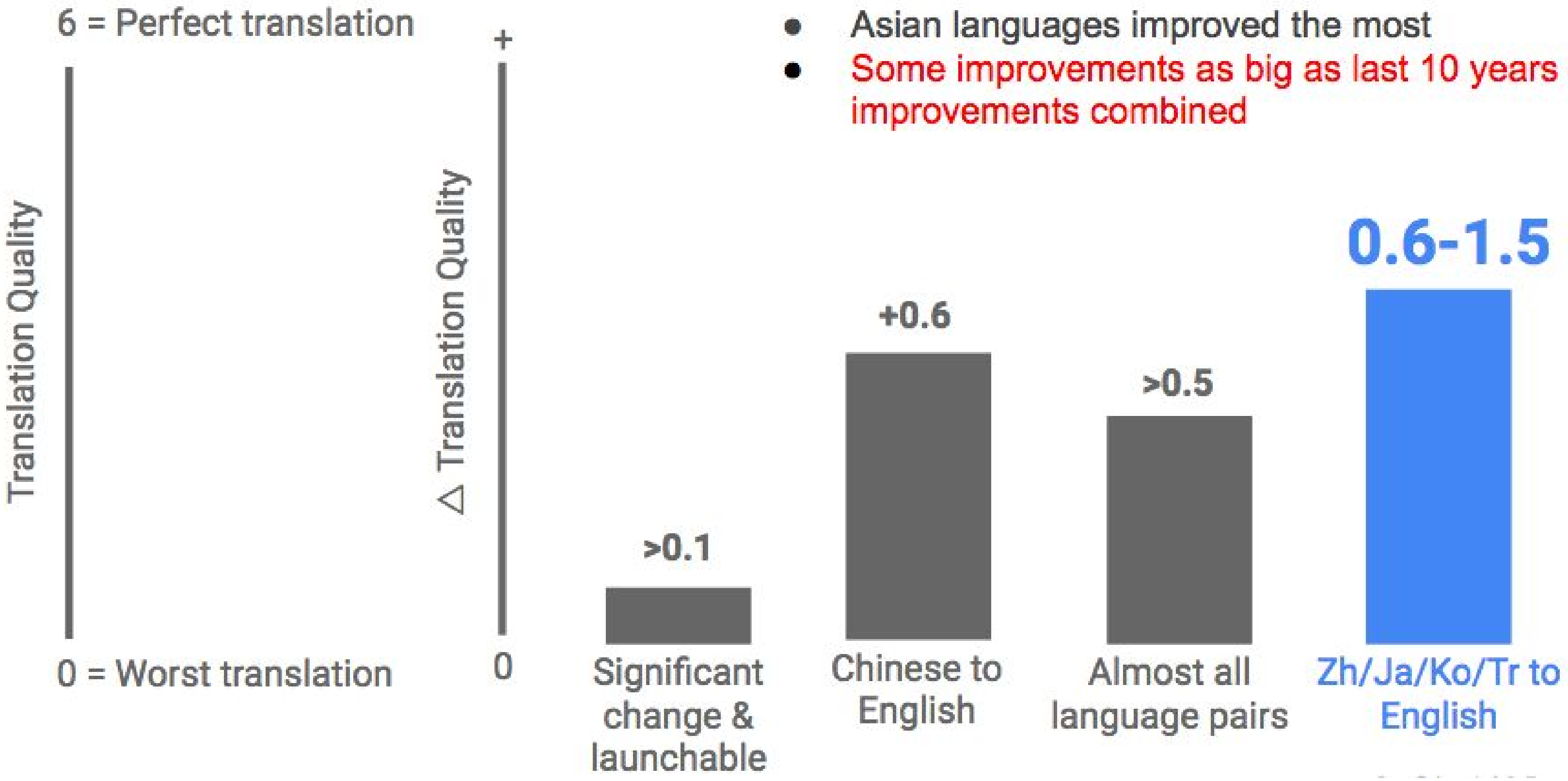
Silent launch in Japan…

4 hours ago	5 hours ago	6 hours ago	7 hours ago
Google翻訳	Google翻訳	紅の豚	#勇者ヨシヒコ
#三四郎ann0	#bananamoon	#勇者ヨシヒコ	紅の豚
#bananamoon	#勇者ヨシヒコ	#NAOMIの部屋	#jojo_anime
#勇者ヨシヒコ	#三四郎ann0	Google翻訳	#Lostorage
#NAOMIの部屋	#NAOMIの部屋	#bananamoon	ラブゲ
#Lostorage	#Lostorage	#Lostorage	#こち星
ラブゲ	ラブゲ	#jojo_anime	#bananamoon
二人セゾン	二人セゾン	ラブゲ	猫の恩返し
Taka	Taka	猫の恩返し	フィオ
予想的中	猫の恩返し	砂の塔	Google翻訳

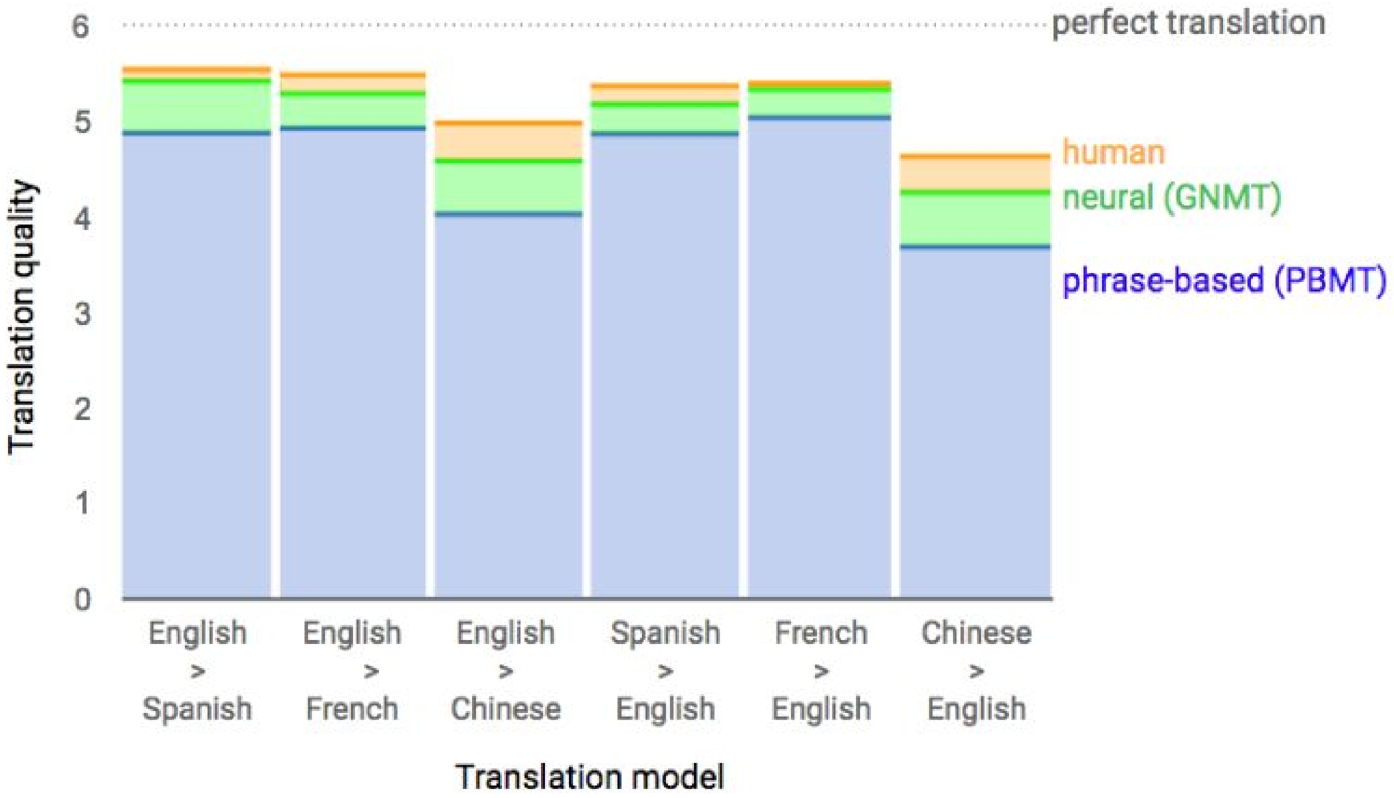
(November 2016)

Quality improvements

- Asian languages improved the most
- Some improvements as big as last 10 years of improvements combined



Relative improvement



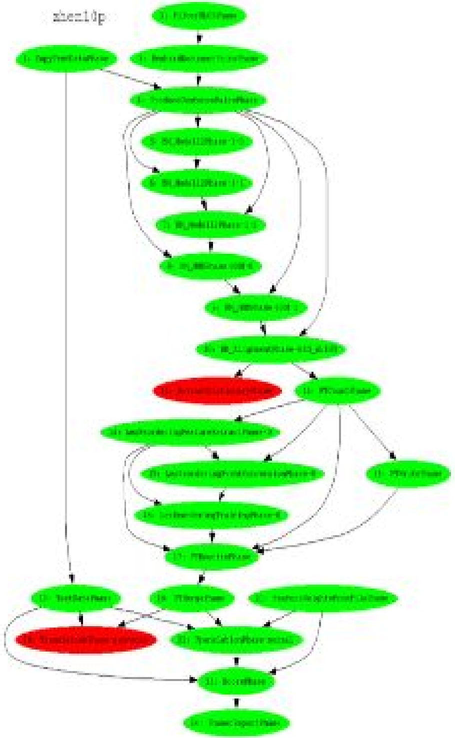
Does quality matter?

+75%

Increase in daily English - Korean translations on Android over the past six months

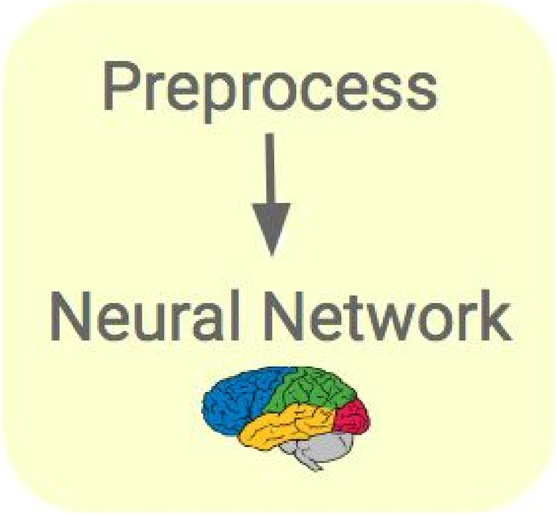
Old: Phrase-based translation

- Lots of individual pieces
- Optimized somewhat independently



New: Neural machine translation

- End-to-end learning
- Simpler architecture
- Plus results are much better!



Outline

Recent Results

Brief History

Word-based translation

Neural translation

What’s next

Brief history of MT

1954: IBM translates 49 Russian sentences with a 250-word dictionary and 6 grammar rules.

Thomas J Watson “I see in this an instrument that will be helpful in working out the problems of world peace...”

Newest Electronic Brain Even Translates Russian

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 instrument 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.”

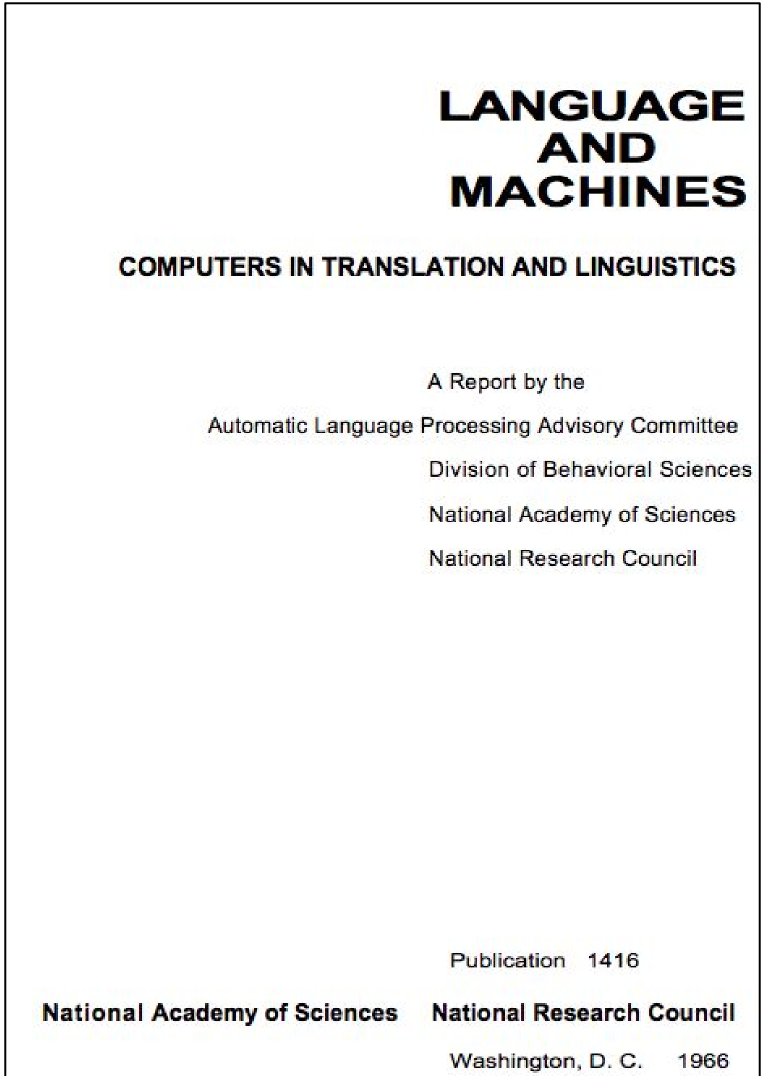
Dostert, who was in charge of installing the original simultaneous translation system at the United Nations, echoed the thought.

Frank James White, recently sentenced in a British stock swindle, once sold $28,000 worth of honey by mail although he had no honey to sell.

Brief history of MT

1966: Alpac publishes a report concluding that years of research haven't produced useful results. Federal funding for MT research dries up… “AI winter”.

Yehoshua Bar Hillel “The unreasonableness of aiming at fully-automatic high-quality translation is stressed...”



LANGUAGE
AND
MACHINES

COMPUTERS IN TRANSLATION AND LINGUISTICS

A Report by the
Automatic Language Processing Advisory Committee
Division of Behavioral Sciences
National Academy of Sciences
National Research Council

Publication 1416

National Academy of Sciences National Research Council

Washington, D. C. 1966

Brief history of MT

1988: IBM Model 1 based on parallel corpora and simple statistical models revives MT.

Peter Brown, Robert Mercer, et al.

The Mathematics of Machine Translation: Parameter Estimation

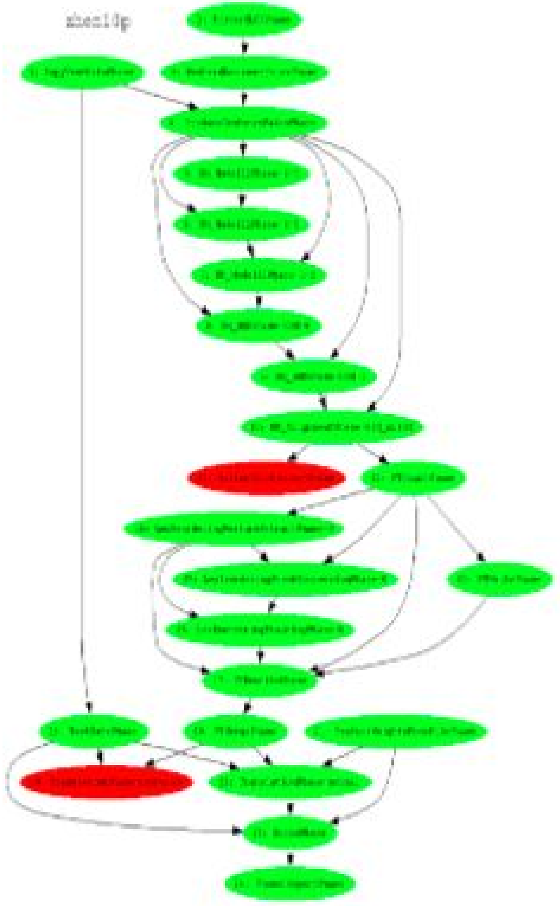
Peter F. Brown, Stephen A. Della Pietra
Vincent J. Della Pietra, Robert L. Mercer

The availability of large, bilingual corpora has stimulated recent interest in algorithms for manipulating them. A number of authors have discussed algorithms for extracting from such corpora pairs of sentences that are translations of one another. In the course of our work on machine translation, we have developed a series of five statistical models of the translation process. Here, we describe these models and show that it is possible to estimate their parameters automatically from a large set of pairs of sentences. We show, further, that it is possible to align the words within pairs of sentences algorithmically. We have a great deal of data in French and English from the proceedings of the Canadian Parliament. For this reason we have restricted our work to these two languages, but we feel that because our algorithms have minimal linguistic content they would work well on other pairs of languages. We also feel, again because of the minimal linguistic content of our algorithms, that it is reasonable to argue that word-by-word alignments are inherent in any sufficiently large bilingual corpora.

Brief history of MT

1995-2014

- Words -> Phrases
- Scaling up
- A lot of tuning



Outline

Recent Results

Brief History

Word-based translation

Neural translation

What’s next

Translation probabilities

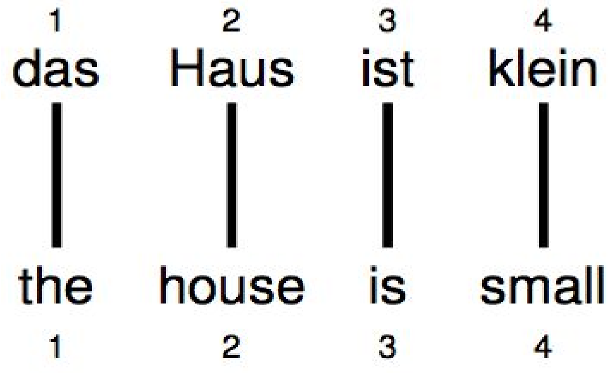
- How to translate a word? Look it up!
  - **Haus** -> house, building, home, household, shell
  - But some more frequent than others…
- We really want to estimate translation probabilities

Haus

e	$t(e\|f)$
house	0.8
building	0.16
home	0.02
household	0.015
shell	0.005

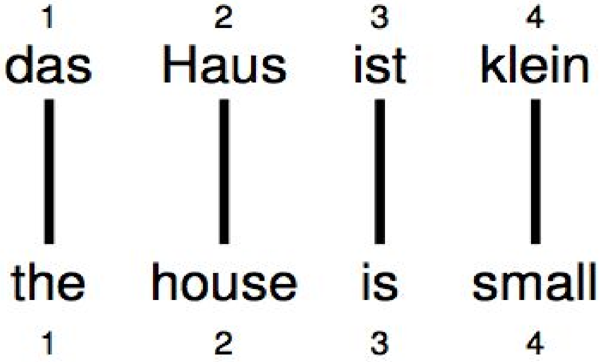
Alignments

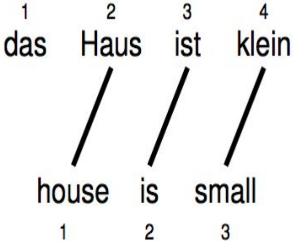
- Supposing we have parallel text...
- If we had alignments, we could count translations -> estimate probabilities

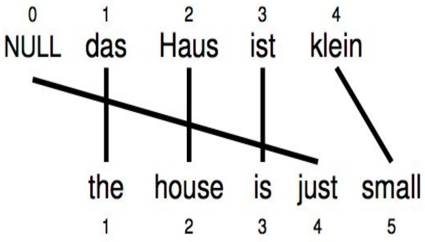


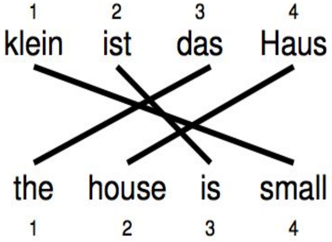
Alignments

- Supposing we have parallel text...
- If we had alignments, we could count translations -> estimate probabilities
- Alignments come in different flavors



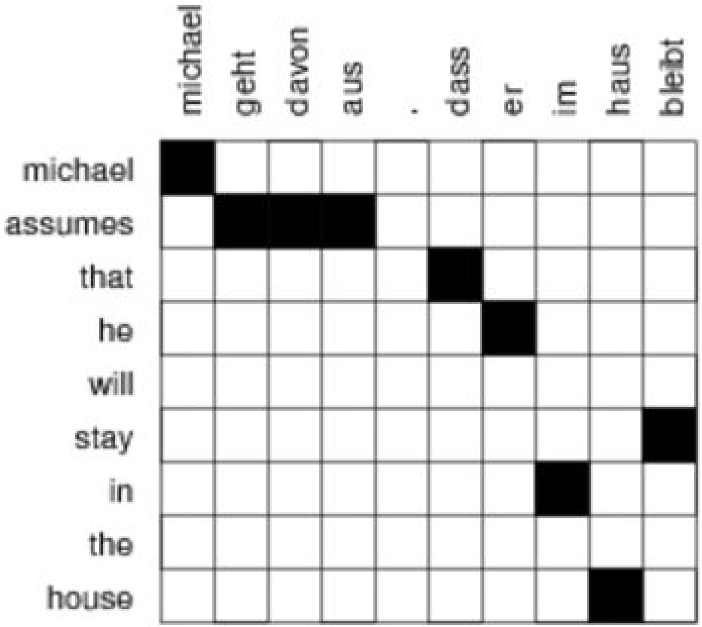




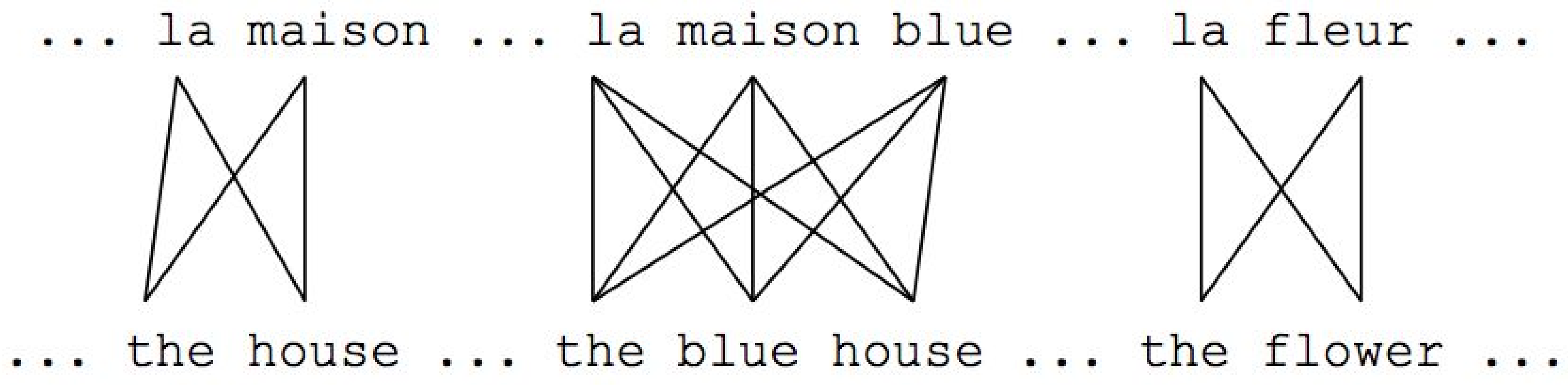


Alignments are latent

- Classic chicken and egg problem!
- If we had alignments, we could estimate translation probabilities...
- ...and if we had translation probabilities, we could generate alignments
- **Solution**: Expectation Maximization Algorithm

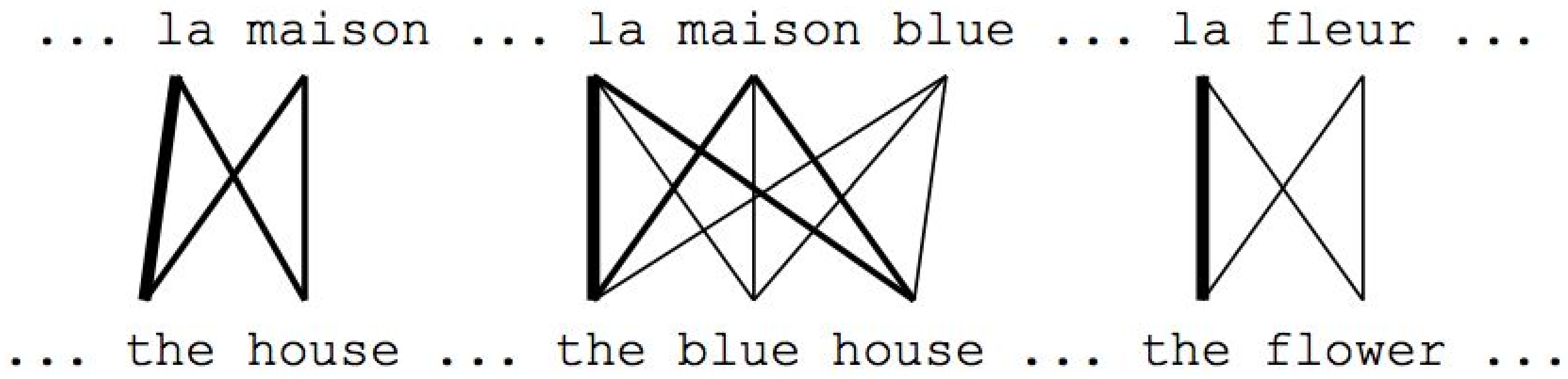


EM for Model 1



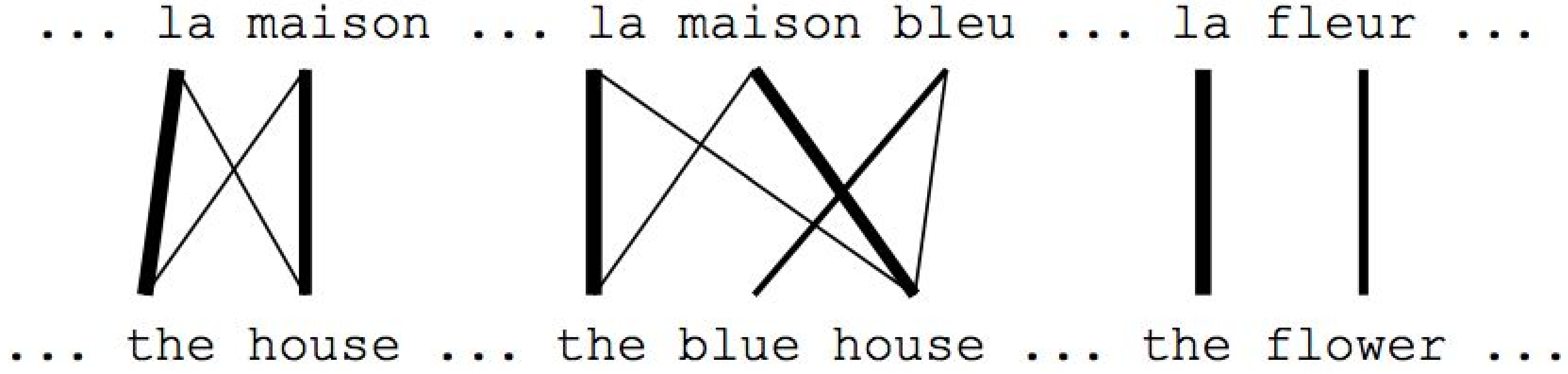
- Initially, assume all alignments are equally likely (E-step 0)
- Estimate translation probabilities using the alignments (M-step 0)

EM for Model 1



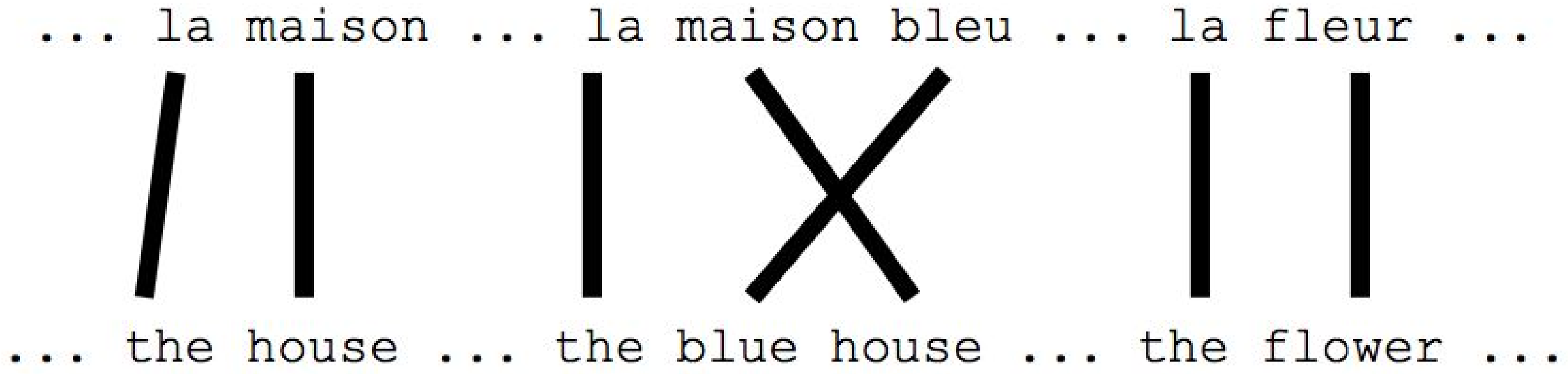
- P(*the*|*la*) > P(*house*|*la*)
- Now, produce new alignments using the updated translation model (E-step 1)
- Again, re-estimate translation model (M-step 1)

EM for Model 1



- After another iteration, *fleur* is aligned to *flower*.

EM for Model 1



- Repeat until convergence
- Note: convergence guaranteed!

p(la|the) = 0.453
p(le|the) = 0.334
p(maison|house) = 0.876
p(bleu|blue) = 0.563
...

Decoding

la maison bleu

?

Decoding

la maison bleu

?

- Use the translation model to generate some hypotheses
  - There are **a lot** of possibilities

TM

```
p(la|the) = 0.453
p(le|the) = 0.334
p(maison|house) = 0.876
p(bleu|blue) = 0.563
...
```

Decoding

- Use the translation model to generate some hypotheses
  - There are **a lot** of possibilities
- Rescore them with a language model
  - This is the **noisy channel** setup

la maison bleu

↓

?

TM

```
p(la|the) = 0.453
p(le|the) = 0.334
p(maison|house) = 0.876
p(bleu|blue) = 0.563
...
```

LM

```
P(house|the)
P(house|blue)
P(blue|the)
...
```

Decoding

- Use the translation model to generate some hypotheses
  - There are **a lot** of possibilities
- Rescore them with a language model
  - This is the **noisy channel** setup
- Beam search
  - Maintain a fixed size stack of partial hypotheses

la maison bleu

?

TM

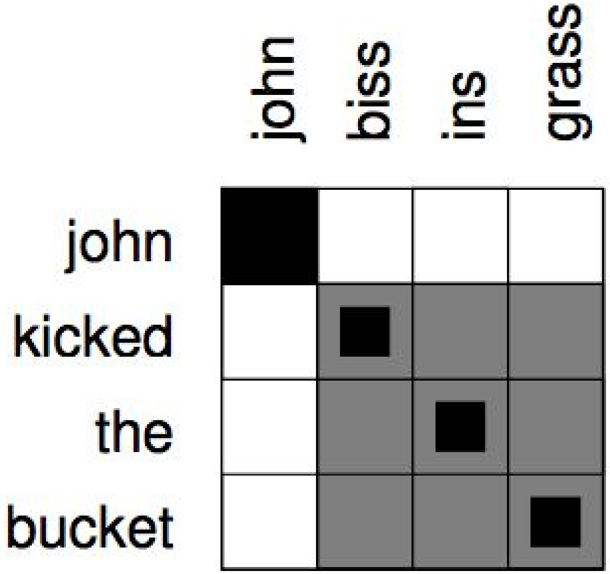
```
p(la|the) = 0.453
p(le|the) = 0.334
p(maison|house) = 0.876
p(bleu|blue) = 0.563
...
```

LM

```
P(house|the)
P(house|blue)
P(blue|the)
...
```

IBM Models -> Phrase translation

- Model 1: Lexical translation
- Model 2: Adds absolute reordering model
- Model 3: Adds fertility model
- Model 4: Relative reordering
- Model 5 (and 6): Fix deficiencies
- Phrase translation



Outline

Recent Results

Brief History

Word-based translation

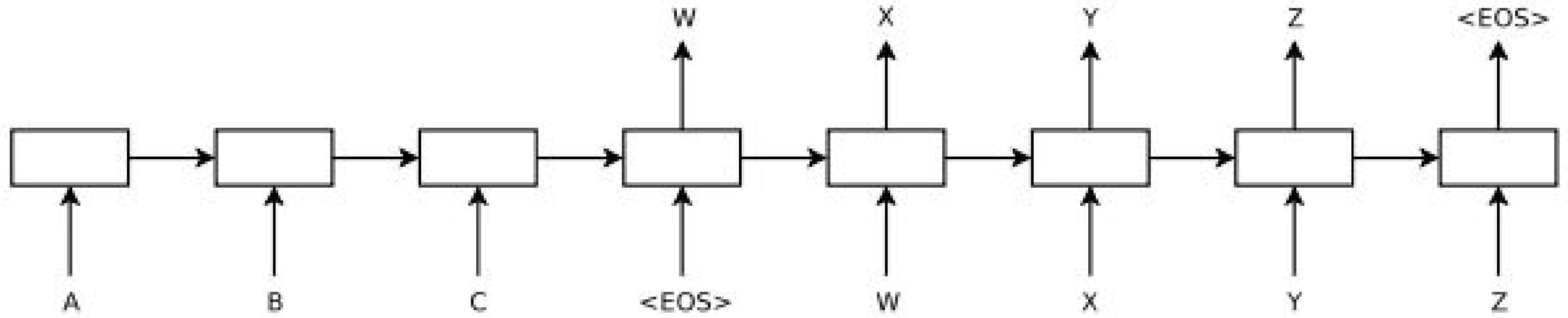
Neural translation

What's next

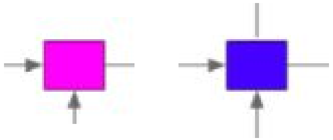
2014: Sequence to Sequence

Sequence to Sequence Learning with Neural Networks -- Sutskever, Vinyals, Le

- No alignments!
- No language model!

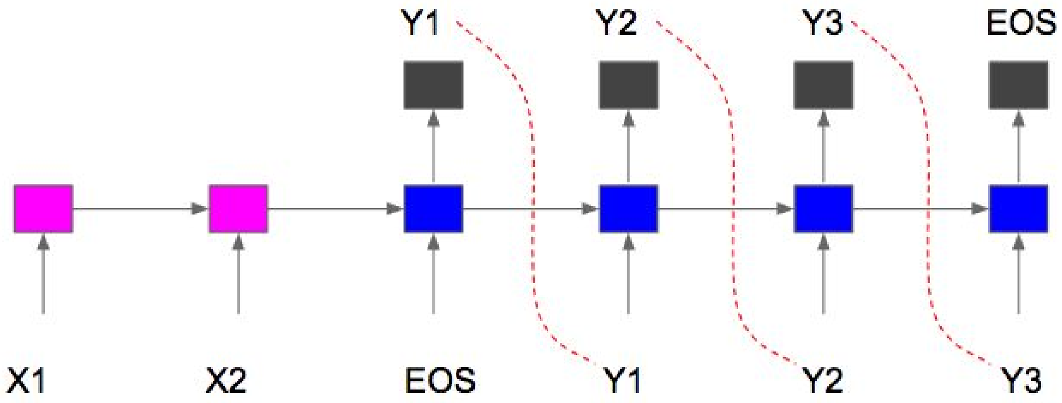


Encoder/Decoder

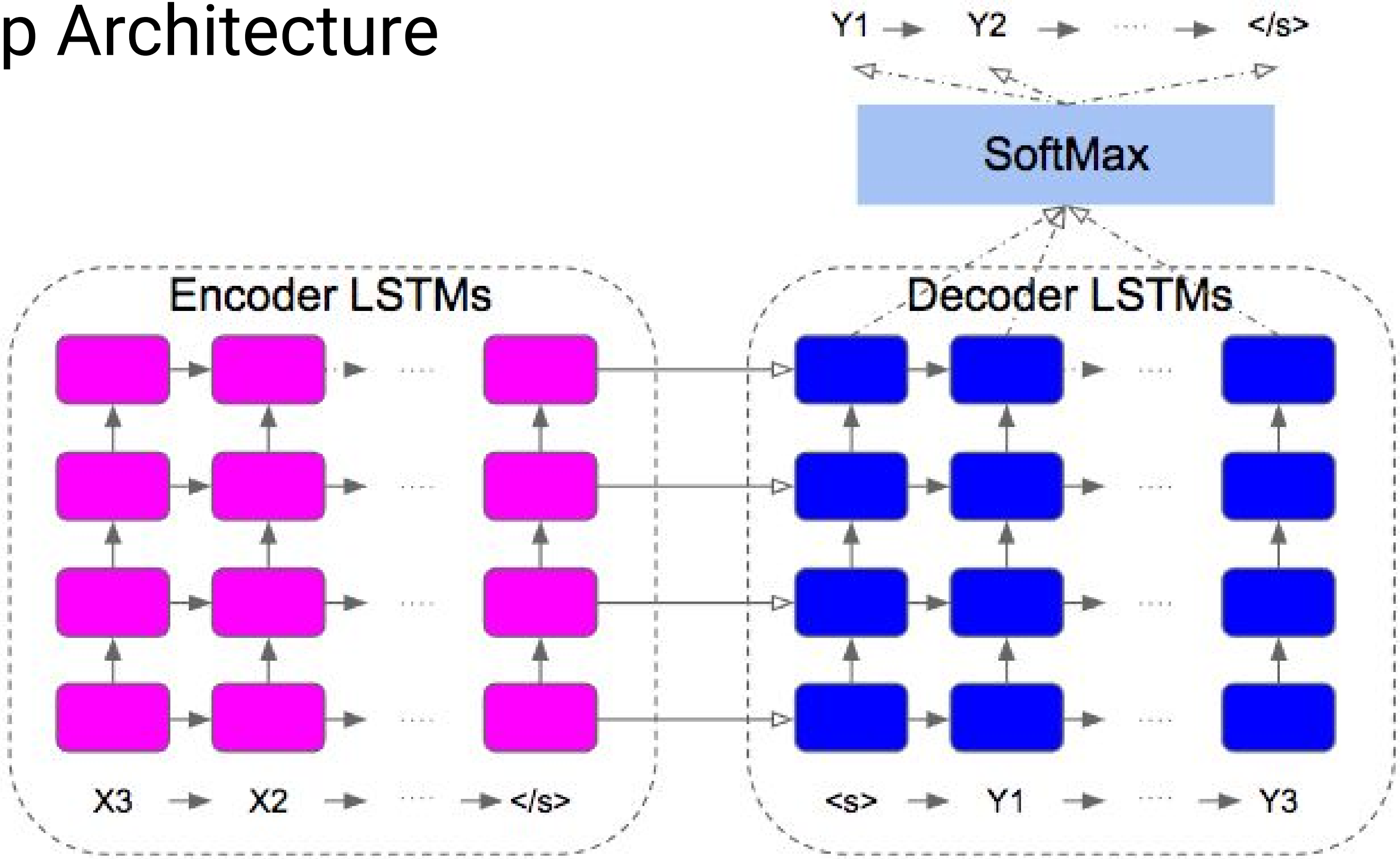


Encoder/Decoder Recurrent Neural Nets

- Learn to map: X1, X2, EOS -> Y1, Y2, Y3, EOS
- In principle, any lengths should work
- RNN -> LSTM



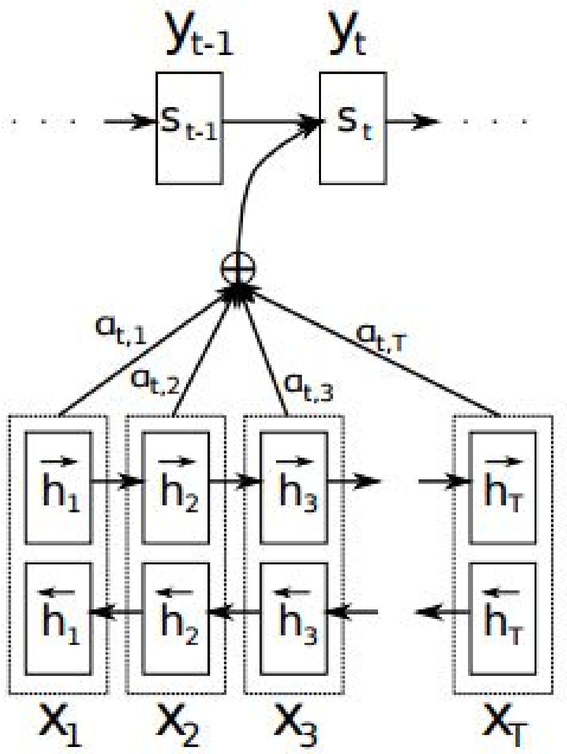
Deep Architecture



2014: Attention Mechanism

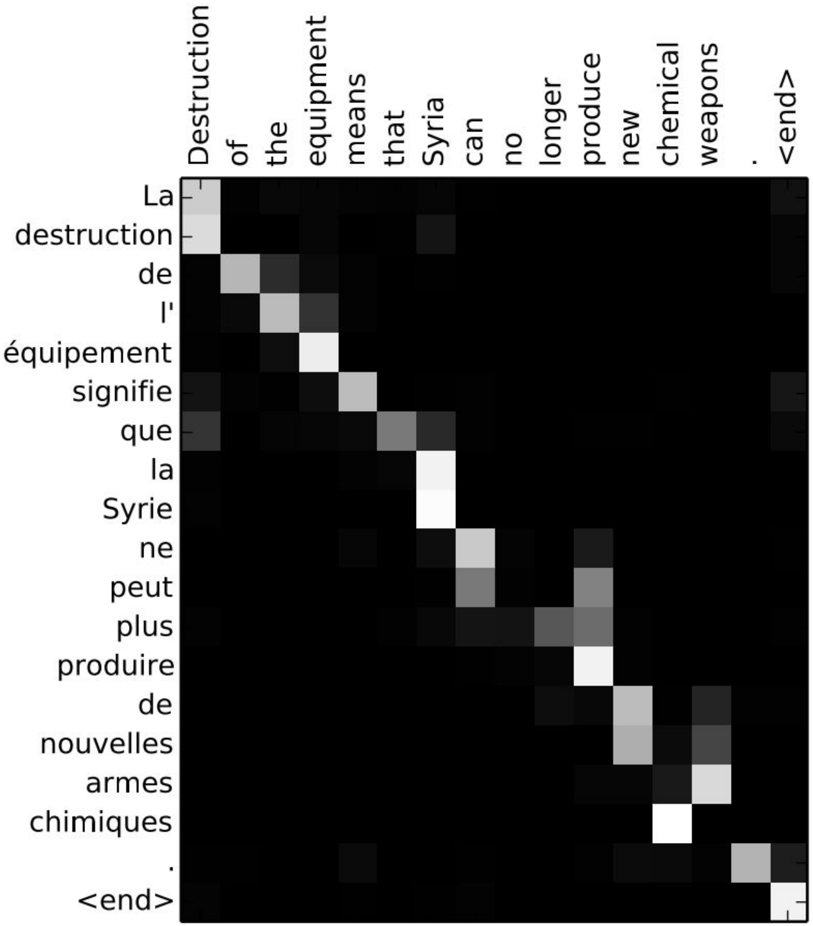
Neural Machine Translation by Jointly Learning to Align and Translate -- Bahdanau, Cho, Bengio

- Give decoder access to all encoder states
- Now quality independent of sentence length



Attention Mechanism

- Also, we can retrieve approximate alignments from the attention weights.



Outline

Recent Results

Brief History

Word-based translation

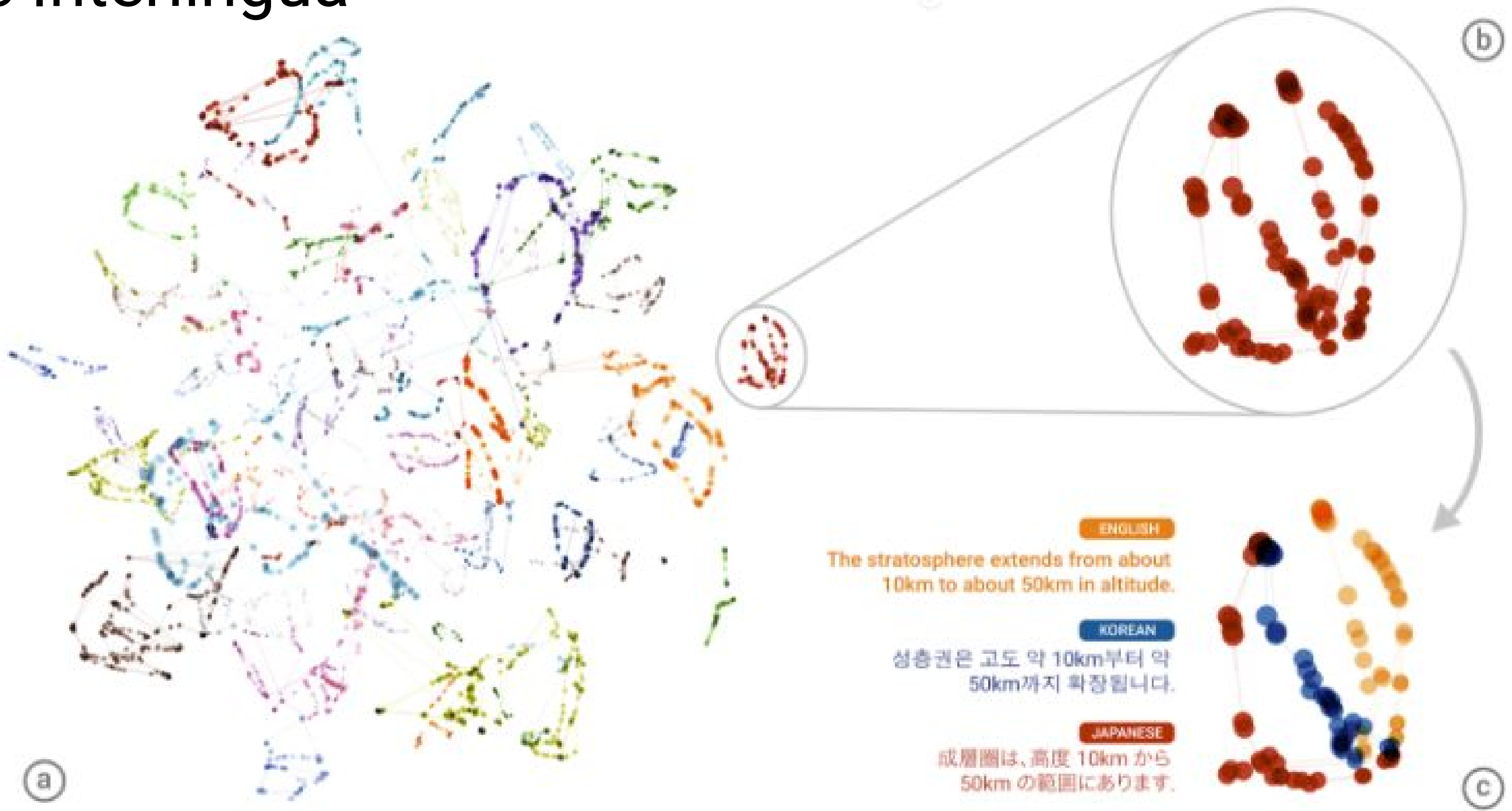
Neural translation

What’s next

Multilingual Models

- Model several language pairs in single model
- Prepend source with additional token to indicate target language
    - Translate to Spanish:
        - <2es> How are you </s> -> Cómo estás </s>
    - Translate to English:
        - <2en> Cómo estás </s> -> How are you </s>
- No other changes to model architecture!

The Interlingua



What’s next?

- Debugging is hard
- Full document translation
- Use more training data
- More efficient models (avoid RNNs entirely?)
- Translation -> other language tasks