Into Language Technologies II

Alberto Barrón-Cedeño
Alma Mater Studiorum - Università di Bologna
a.barron@unibo.it
@_albarron_

Session 2
21st December 2022

Disclaimer

I have reduced the math to the minimum necessary
to show the basics of MT
(but we will see some equations)

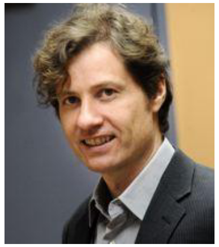
Second Disclaimer

Materials partially derived from lectures and notes by...


Hassan Sajjad
Dalhousie University


Fahim Dalvi
QCRI


Philipp Koehn
Johns Hopkins


Michael Collins
Columbia/Google


Fabienne Cap
Artificial Solutions

https://hsajjad.github.io/pages/dl4mt/
http://mt-class.org
http://www.cs.columbia.edu/~cs4705/

Overview

- **Introduction:** a very brief “history” of MT
- **The basic *ingredients* of MT**: mostly (late) 1990s
  - Translation Model
  - Language Model
  - Decoder
  - Evaluation
- **Today:** only 1 slide!

The Inception of MT

When I look at an article in Russian, I say: “This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode”.

Warren Weaver, 1947 (Matematician)

Introduction

“Pre-History”: Rule-based Systems

- Systran (1968)   ← **used by Google until 2007**
- Canada’s Météo system for weather forecasts (1976)
- Logos and Metal (1980s)

“Pre-History”: Rule-based Systems

Google’s announcement on switching to their in-house (statistical) model: 25 pairs
https://googlesystem.blogspot.com/2007/10/google-translate-switches-to-googles.html

“I tried to translate an english website into french and it's **still not a great translation**.”

“I typed in some basic phrases from Eng>French and was surprised at **how poorly they were translated**. Google translate has a long way to go before matching babelfish.”

“You shouldn't expect perfection from an automatic translation. **Humans are always better for this job, but that requires time and money**.”

“History”: Statistical Machine Translation

- IBM (1990s): Models 1-5
- Moses and Google (mid 2000s): Phrase-based MT
- Around 2010: commercial viability

“Present”: Neural Machine Translation

- First neural models (mid 2010s)
- New state of the art (2016)
- People are still trying to find better ways to *drive* NMT

The basic ingredients of MT

What is involved in MT?

- **Translation model** assigns a conditional probability $p(f \mid e)$ to a pair of sentences
    - It learns word- and phrase-level translations
- **Language model** assigns a probability $p(e)$ for sentence *e*
    - It learns to generate fluent translations
- **Decoder** generates a translation given an input text
    - It produces a translation from the *trained* translation and language models

Let’s play Arrival

The Sith-English parallel training corpus

(exercise designed by Fabienne Cap)

Let’s play Arrival: the Sith-English parallel corpus

Tegu mus minti kait mes itik kash .
Let ’s see how we get in .

Tave dury kia tave sodas kash artija .
The door to the garden is closed .

Kad kait isar itik mes kash ?
But how do we get in ?"

tave	→	the
isar	→	do
kait	→	how
kait mes itik kash	→	how do we get in
Kait isar itik mes kash	→	how do we get in
.	→	.
?	→	?

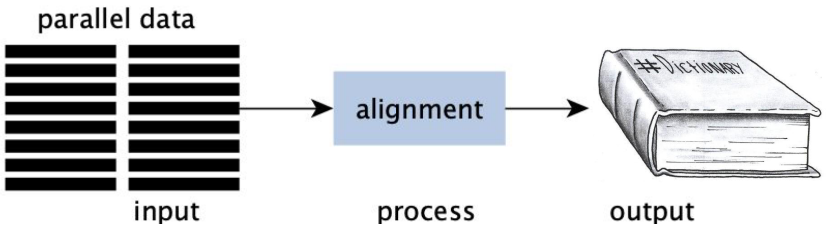
What do we get those pairs?

From a set of parallel sentences, we can

- learn a dictionary
- find ambiguous words
- one to many and many to one translations

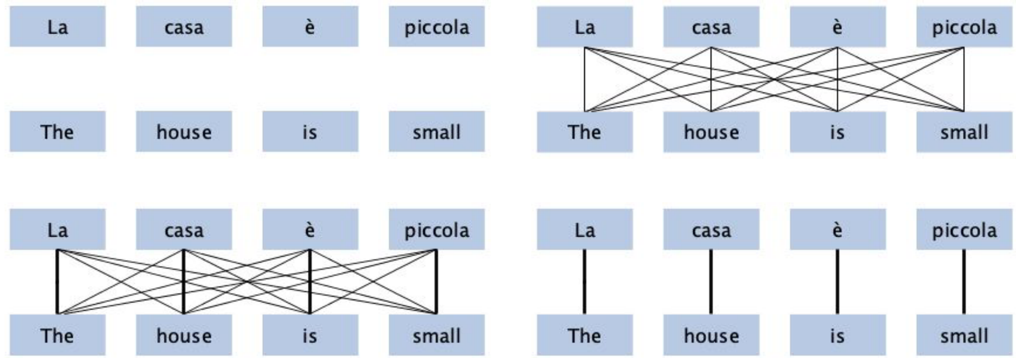
This is what machine translation needs!

Translation Model



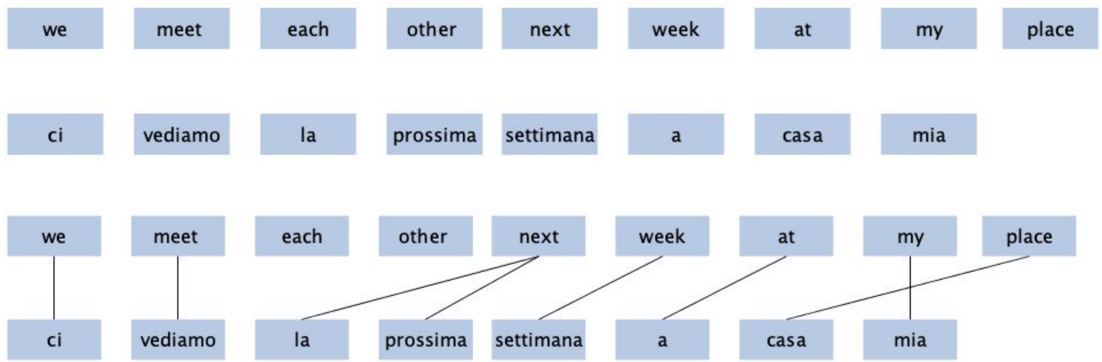
Translation Model

Alignment



Translation Model

Alignment (a more realistic example)



Translation Model

Dictionary

la		casa		è		piccola	
e	*t(e|f)*	*e*	*t(e|f)*	*e*	*t(e|f)*	*e*	*t(e|f)*
the	0.700	house	0.800	is	0.800	small	0.400
that	0.150	home	0.160	’s	0.160	little	0.400
which	0.075	building	0.020	exists	0.020	short	0.100
who	0.050	household	0.015	has	0.015	minor	0.060
this	0.025	shell	0.005	are	0.005	petty	0.040

(adapted from Koehn’s en-de example)

Let us see how to produce a Sith-English dictionary *for real*

What is an MT model?

- Translation model assigns a conditional probability $p(f \mid e)$ to a pair of sentences
  - learn word-level and phrase-level translations
- **Language model assigns a probability** $p(e)$ **for sentence *e***
  - learn to generate fluent translations
- Decoder
  - translation generation component
  - produce a translation from the trained translation and language models

Language Model

Let’s try again

... a ________ ...

car	cars	water	cat

Language Model

It must be recognized that the notion “probability of a sentence” is an entirely useless one, under any known interpretation of this term.

Noam Chomsky, 1969

Language Model

Let’s try again

... a ________ ...

car and *cat* both work

car	cars	water	cat

Can you fill the gaps?

Forlì è un comune **italiano** di 117 627 abitanti, capoluogo **della** provincia di Forlì-Cesena in Romagna. È sede vescovile della diocesi **di** Forlì-Bertinoro.

https://it.wikipedia.org/wiki/Forlì

Language Model

- Reordering

  p_{LM}(the house is small) > p_{LM}(small is the house)

- Word choice

  p_{LM}(I am going home) > p_{LM}(I am going house)

Language Model

Let's try again

... a _________ ...

car and *cat* both work

car | cars | water | cat

John is driving a _________ ...

Language Model

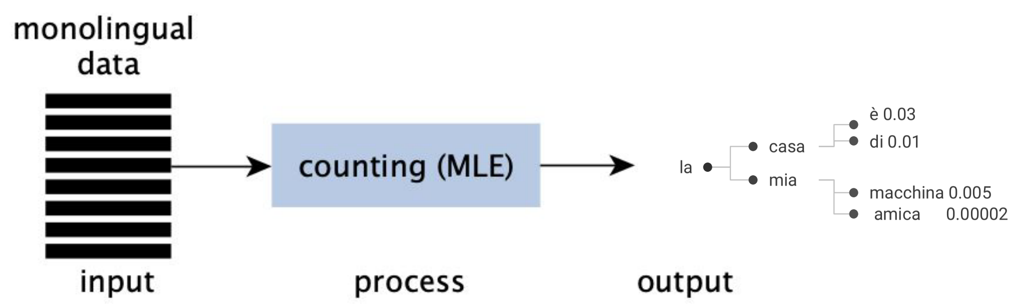
Given sequence $W = w_1, w_2, w_3, \ldots, w_n$, what is $p(W)$?

By the chain rule…

$$p(w_1, w_2, w_3, \ldots, w_n) = p(w_1)p(w_2 \mid w_1)p(w_3 \mid w_1, w_2) \ldots p(w_n \mid w_1, w_2, \ldots, w_{n-1})$$

Do you spot any issue?

Language Model



Language Model

Markov assumption

- only a limited previous history matters, as the further you go in the past, the less relevant the information becomes

- *k*-th order Markov model; here with *k=2*

$$p(w_1, w_2, w_3, \ldots, w_n) \cong p(w_1)p(w_2 \mid w_1)p(w_3 \mid w_2) \ \cdots \ p(w_n \mid w_{n-1})$$

What is involved in MT?

- Translation model assigns a conditional probability $p(f \mid e)$ to a pair of sentences
  - It learns word-level and phrase-level translations
- Language model assigns a probability $p(e)$ for sentence *e*
  - It learns to generate fluent translations
- **Decoder generates a translation given an input text**
  - It produces a translation from the trained translation and language models

Full Model

IBM Model 1: translation probability

- Given a foreign sentence $\mathrm{f} = (f_1, \ldots, f_{|\mathrm{f}|})$ of length *|f|*
- and an English sentence $\mathrm{e} = (e_1, \ldots, e_{|\mathrm{e}|})$ of length *|e|*
- with an alignment of each English word e_j to a foreign word f_i according to the alignment function $a : j \to i$

$$p(\mathrm{e}, a \mid \mathrm{f}) = \frac{\epsilon}{(|\mathrm{f}|+1)^{|\mathrm{e}|}} \prod_{j=1}^{|\mathrm{e}|} t(e_j \mid f_{a(j)})$$

where ϵ is a normalisation constant

(adapted from Koehn's en-de example)

Full Model

IBM Model 1: translation probability

$$p(\mathrm{e}, a \mid \mathrm{f}) = \frac{\epsilon}{(|\mathrm{f}|+1)^{|\mathrm{e}|}} \prod_{j=1}^{|\mathrm{e}|} t(e_j \mid f_{a(j)})$$

la		casa		è		piccola	
e	*t(e|f)*	*e*	*t(e|f)*	*e*	*t(e|f)*	*e*	*t(e|f)*
the	0.700	house	0.800	is	0.800	small	0.400

$$p(\mathrm{e}, a \mid \mathrm{f}) = \frac{\epsilon}{(4+1)^4} \times t(\mathrm{the} \mid \mathrm{la}) \times t(\mathrm{house} \mid \mathrm{casa}) \times t(\mathrm{is} \mid \mathrm{è}) \times t(\mathrm{small} \mid \mathrm{piccola})$$

$$p(\mathrm{e}, a \mid \mathrm{f}) = \frac{\epsilon}{5^4} \times 0.7 \times 0.8 \times 0.8 \times 0.4$$

$$p(\mathrm{e}, a \mid \mathrm{f}) = 0.00028672\ \epsilon$$

(adapted from Koehn's en-de example)

Decoder

Produce (search for) a translation from a trained translation model and language model

Er fährt sehr schnell

$\downarrow$ How?

He drives very fast

(example borrowed from Sajjad & Dalvi)

Decoding (Translation generation)

Er | fährt | sehr | schnell

Translation is generated from left to right

Decoding (Translation generation)

Er | fährt | sehr | schnell

he

it

Look at the source sequence to get choices for the first target word

Decoding (Translation generation)

Er | fährt | sehr | schnell

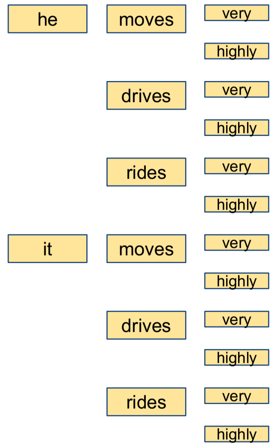
he → moves | drives | rides

it → moves | drives | rides

Get all the potential choices for the second target word

Decoding (Translation generation)

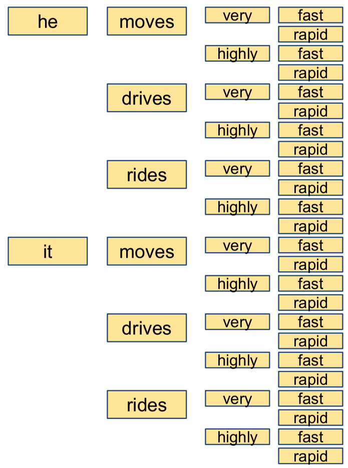
Er fährt sehr schnell



Continue until the full sentence is generated

Decoding (Translation generation)

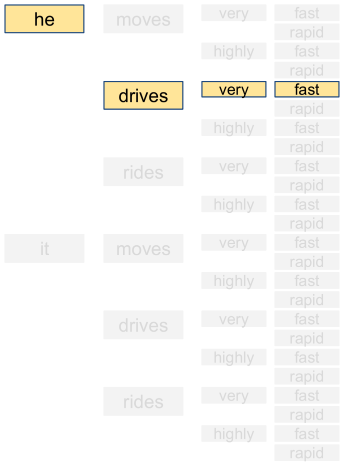
Er fährt sehr schnell



Continue until entire sentence is generated

Decoding (Translation generation)

Er fährt sehr schnell



Rank all possible choices and select the best translation

Could you translate en-si?

Evaluation

Five translations of a Chinese sentence:

这个机场的安全由以色列方面负责.

Israeli officials are responsible for airport security.

Israel is in charge of the security at this airport.

The security work for this airport is the responsibility of the Israeli government.

Israeli side was in charge of the security of this airport.

Israel is responsible for the airport’s security.

(borrowed from Koehn)

Evaluation

How good is a translation in terms of meaning and fluency?

- **Human evaluation**
  - Adequacy. The translation holds the meaning of the source sentence?
  - Fluency. Is it a grammatically and syntactically fluent sentence?
- **Automatic evaluation**
  - BLEU; Meteor; TER; WER; Bertscore… and many, many, many more
- **Quality estimation**

Automatic Evaluation - BLEU

- Considers two aspects between a translation output and a reference
  - *n*-gram overlap, with *n=[1, 4]*
  - brevity penalty (length difference)

$$\text{BLEU} = \min\left(1, \frac{\text{output length}}{\text{reference length}}\right)\left(\prod_{i=1}^{4} \text{precision}_i\right)^{\frac{1}{4}}$$

Automatic Evaluation - BLEU

Reference	Israeli officials are responsible for airport security
Output	Israeli officials are responsible for security

Automatic Evaluation - BLEU

Reference	Israeli officials are responsible for airport security
Output	Israeli officials are responsible for security

Today-*ish*

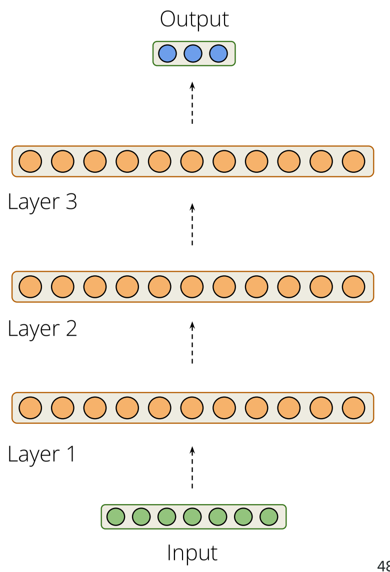
Automatic Evaluation - BLEU

Reference	Israeli officials are responsible for airport security
Output	Israeli officials are responsible for security

4-gram matches *1*-gram match

Deep Machine Translation

- State of the art
- Multilingual translation
- Zero-shot translation
- Neuron manipulation to change the outcome
  - masculine → feminine
  - passive → active
  - singular → plural
- Understanding what is going on!



Automatic Evaluation - BLEU

Reference	Israeli officials are responsible for airport security
Output	Israeli officials are responsible for security

$$\mathrm{BLEU} = \min\left(1, \frac{\text{output length}}{\text{reference length}}\right)\left(\prod_{i=1}^{4} \text{precision}_i\right)^{\frac{1}{4}}$$

Metric	Output
1-gram precision	6/6
2-gram precision	4/5
3-gram precision	3/4
4-gram precision	2/3
brevity penalty	6/7
BLEU	68

Enjoy the holidays!

See you in January