

# NLP Machine Translation

Sunil Murthy, Yawen Zhang, Tetsumichi Umada

# Quick Facts

- EU spend more than € 1.1 billion translation costs each year.
  - => MT will save a lot of money
- Google Translate
  - 103 languages are supported
    - When GT was launched, it's only English and Arabic
  - more than 100 billion words per day



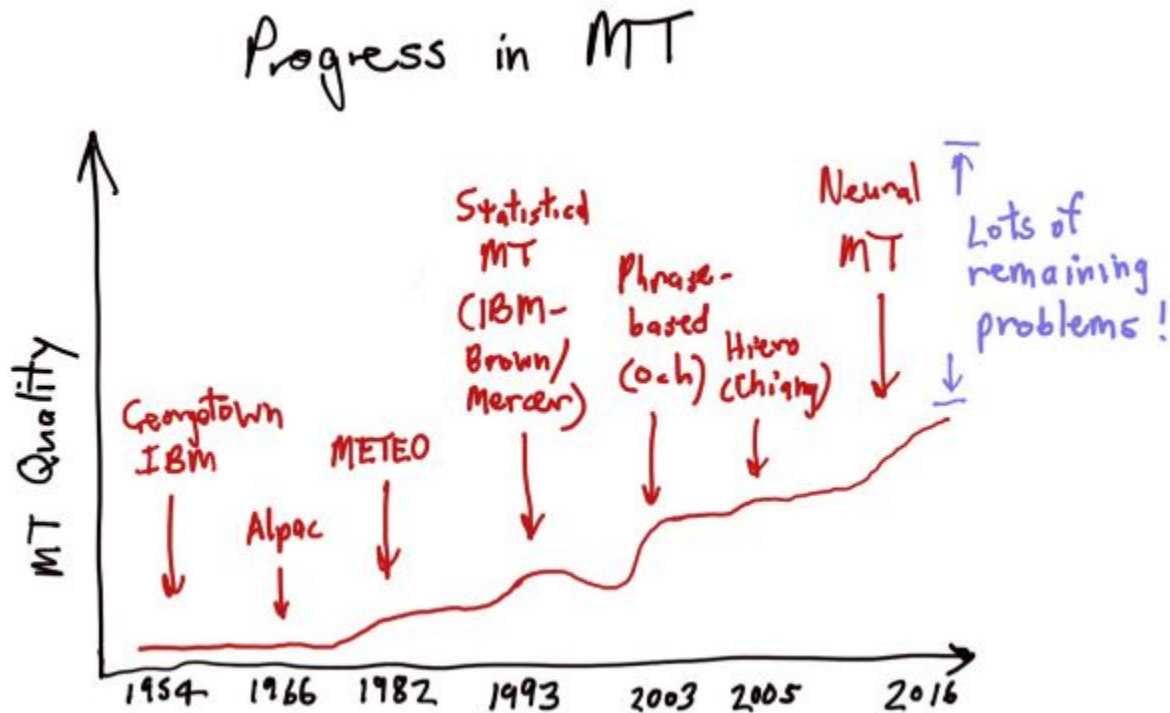
# Why MT is hard?

- Different word order
  - English word order is subject-verb-object (IBM developed Watson)
  - Japanese order is subject-object-verb (IBM Watson developed)
- Word sense
- Tense
  - Spanish: definite time in past vs. unknown time in past
- Pronouns
  - Spanish: -o = I; -as = you; **-a = he/she/it**; -amos = we; -an = they
- Idioms
  - “To kick the bucket” -> “to die”



# Quick History of MT

- IBM: a basic word-for-word translation system in 1954 (Russian to English)
- Statistical machine translation started in 1980s.



Christopher D. Manning

# Approaches

- Statistical based machine translation
  - Use a statistical model developed from analysis of bilingual text corpora
- Rule based translation
  - Translate based on grammars (covering semantic, syntactic, and morphological)
- Hybrid translation
  - Combine statistical and rule based approaches
- Neural machine translation (Google translate)



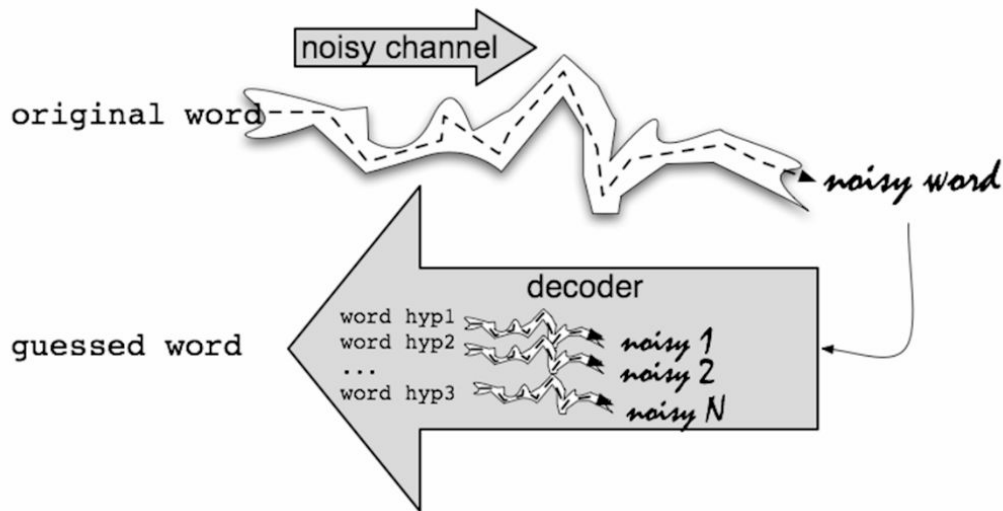
# SMT - Terminology

- IBM Translation models - late 1980s / early 1990s
- Target language (“English”, ...)
  - English sentence,  $\mathbf{e} = (e_1, \dots, e_{l_e})$  of length  $l_e$
- Source language (“German”, “French”, ...)
  - Foreign sentence,  $\mathbf{f} = (f_1, \dots, f_{l_f})$  of length  $l_f$
- Training data
  - Canadian parliamentary proceedings, the **Hansards**.
  - EU parliamentary proceedings, the **Europarl data**.
  - Parallel text

$$(f^{(k)}, e^{(k)}) \text{ for } k = 1 \dots n$$

# Noisy Channel Model

- Language Model -  $P(e)$
- Translation Model -  $P(f | e)$

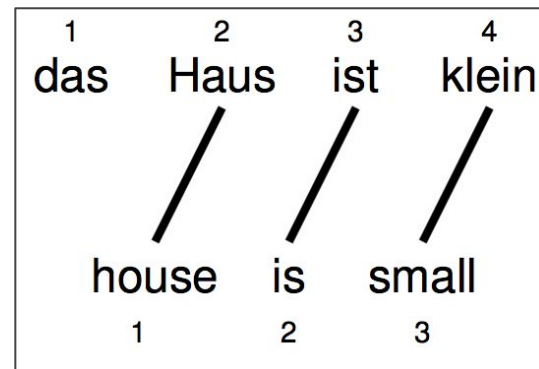
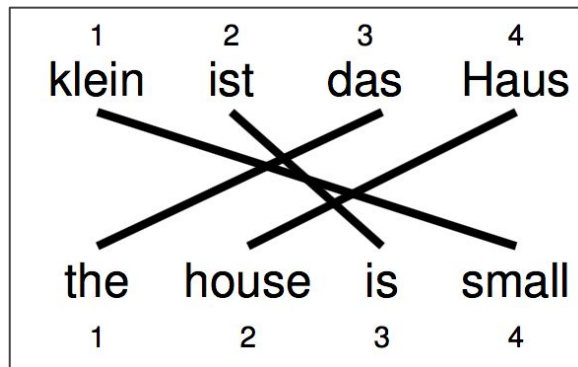
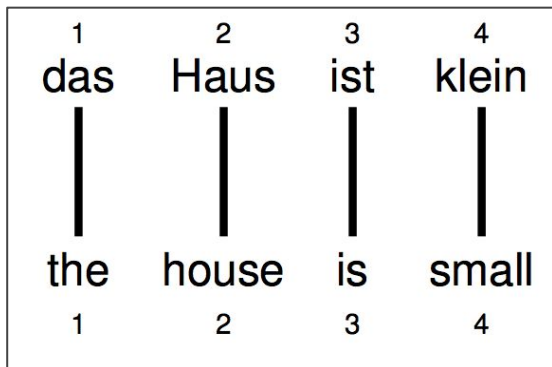


$$e^* = \arg \max_{e \in E} p(e) \times p(f|e)$$

# Word Alignment

- Mapping a target word (english) at position  $i$  to a source word (german) at position  $j$ ,

With a function  $a : i \rightarrow j$







# IBM Translation Models

# IBM Model-1

- Generative model
- Uses lexical translation
- Translation Probability

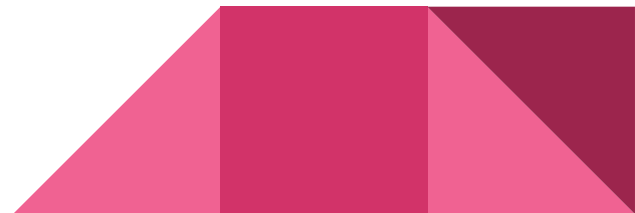
$\mathbf{f}$  = Foreign sentence.

$\mathbf{e}$  = target sentence.

$\mathbf{a}$  = alignment function.

$\epsilon$  = normalization constant.

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



# IBM Model-1 (cont)

- How to estimate translation probabilities,  $t(e|f)$
- Corpora just sentence aligned not word-aligned.
- Problem of incomplete data
- Chicken & egg problem
  - If we had alignments
    - we could estimate model parameters
  - If we had model parameters
    - we could estimate the alignments



# IBM Model-1 (cont.)

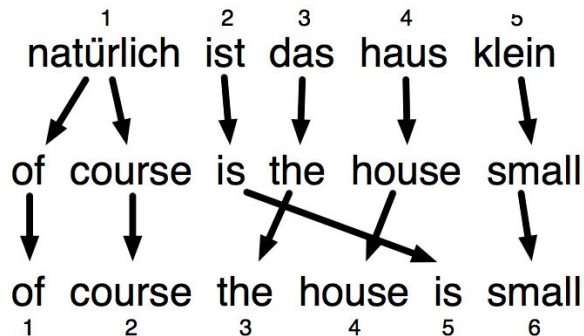
- EM Algorithm for rescue...
  - a. Initialize the model parameters (eg. uniform, randomly...)
  - b. Assign probabilities to missing data
  - c. Estimate model parameters from complete data
  - d. Iterate steps (a) - (c) until model converges.

... la maison ... la maison bleu ... la fleur ...  
/ | | X | |  
... the house ... the blue house ... the flower ...

# IBM Model-2

- Adds absolute alignment to model-1.
- Alignment probability distribution,  $a(i|j, l_e, l_f)$
- Putting everything together,

$$p(\mathbf{e}, \mathbf{a}|\mathbf{f}) = \epsilon \prod_{j=1}^{l_e} t(e_j|f_{a(j)}) a(a(j)|j, l_e, l_f)$$



lexical translation step

alignment step

# IBM Model-3

- Adds fertility model
- Modeled by probability distribution,

$n(\phi|f)$ , in which  $\phi = 0, 1, 2, \dots$

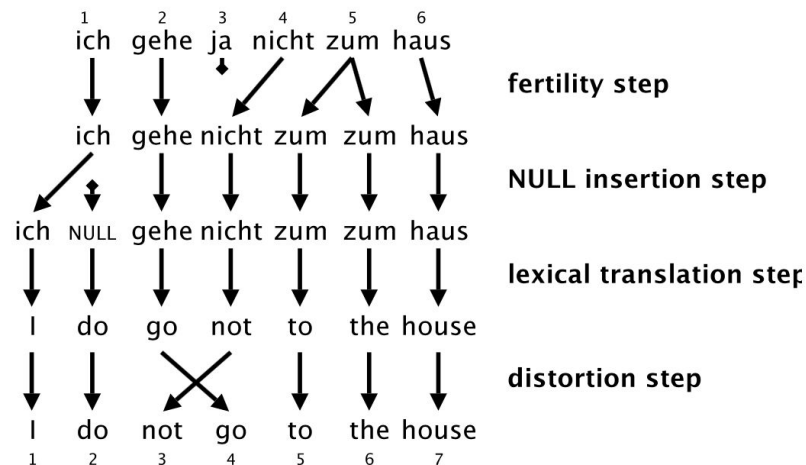
- $$p(\mathbf{e}|\mathbf{f}) = \sum_a p(\mathbf{e}, \mathbf{a}|\mathbf{f})$$
  

$$= \sum_{a(1)=0}^{l_f} \dots \sum_{a(l_e)=0}^{l_f} \binom{l_e - \phi_0}{\phi_0} p_1^{\phi_0} p_0^{l_e - 2\phi_0}$$
  

$$\times \prod_{j=1}^{l_f} \phi_j! n(\phi_j|f_j)$$
  

$$\times \prod_{j=1}^{l_e} t(e_j|f_{a(j)}) d(j|a(j), l_e, l_f)$$

Adding a model of fertility

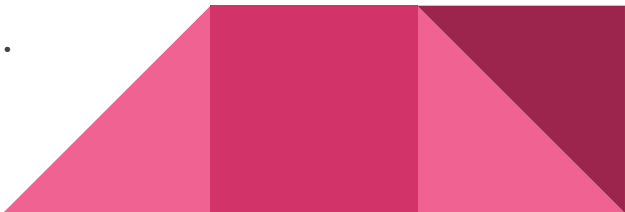


# IBM Model 4 & 5

## Model 4:

- Absolute position for distortion feels wrong.
- Words do not move independent.
- Some words tend to move and some not.

## Model 5:

- Models 1-4 are deficient.
  - Assigns probabilities to impossible translation.
  - Model fixes deficiency by keeping track of vacancies.
- 

# Summary of IBM Model 1-5

Models with increasing complexity

Higher models include more information

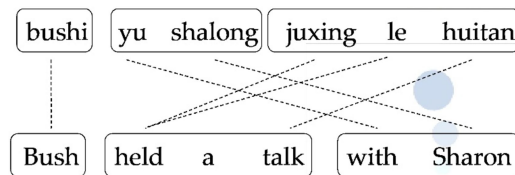
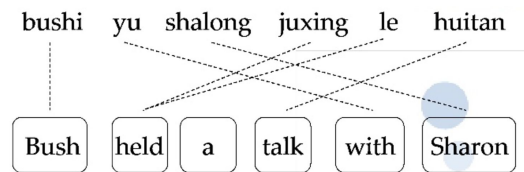
- **IBM Model 1** lexical translation
- **IBM Model 2** adds absolute alignment model
- **IBM Model 3** adds fertility model
- **IBM Model 4** relative alignment model
- **IBM Model 5** fixes deficiency





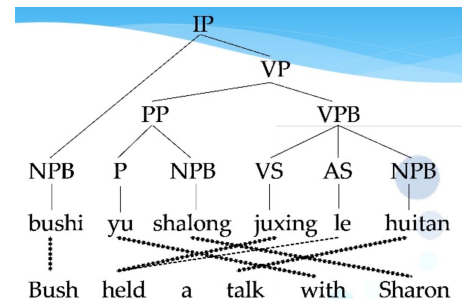
# Progress in MT

# Word, *Phrase*, Syntax, Semantic-based SMT

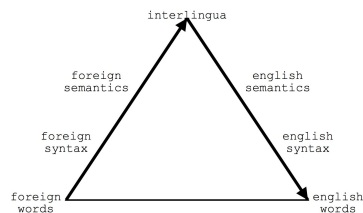


Source	Target	Probability
Bushi (布什)	Bush	0.7
	President	0.2
	US	0.1
yu (与)	and	0.6
	with	0.4
juxing (举行)	hold	0.7
	had	0.3
le (了)	hold	0.01
...	...	...

Source	Target	Probability
Bushi (布什)	Bush	0.5
	president Bush	0.3
	the US president	0.2
Bushi yu (布什与)	Bush and	0.8
	the president and	0.2
yu Shalong (与沙龙)	and Shalong	0.6
	with Shalong	0.4
juxing le huitan (举行了会谈)	hold a meeting	0.7
	had a meeting	0.3
...	...	...



Source	Target	Probability
VPB(VS(juxing) AS(le) NPB(huitan)) (举行了会谈)	hold a meeting	0.6
	have a meeting	0.3
	have a talk	0.1
VPB(VS(juxing) AS(le) x1:NPB) (举行了x1)	hold a x1	0.5
	have a x1	0.5
VP(PP(P(yu) x1:NPB) x2:VPB) (与 x1 x2)	x2 with x1	0.9
IP(x1:NPB VP(x2:PP x3:VPB))	x1 x3 x2	0.7



Language model being defined !

# Pros and Cons about Traditional SMT

- **Pros:**

- System can be built in **a short time**
- Making use of many parallel corpora in machine-readable format
- Generally, not tailored to any specific pair of languages, **broad coverage of language**
- Popular in practice, with just **a few layers**

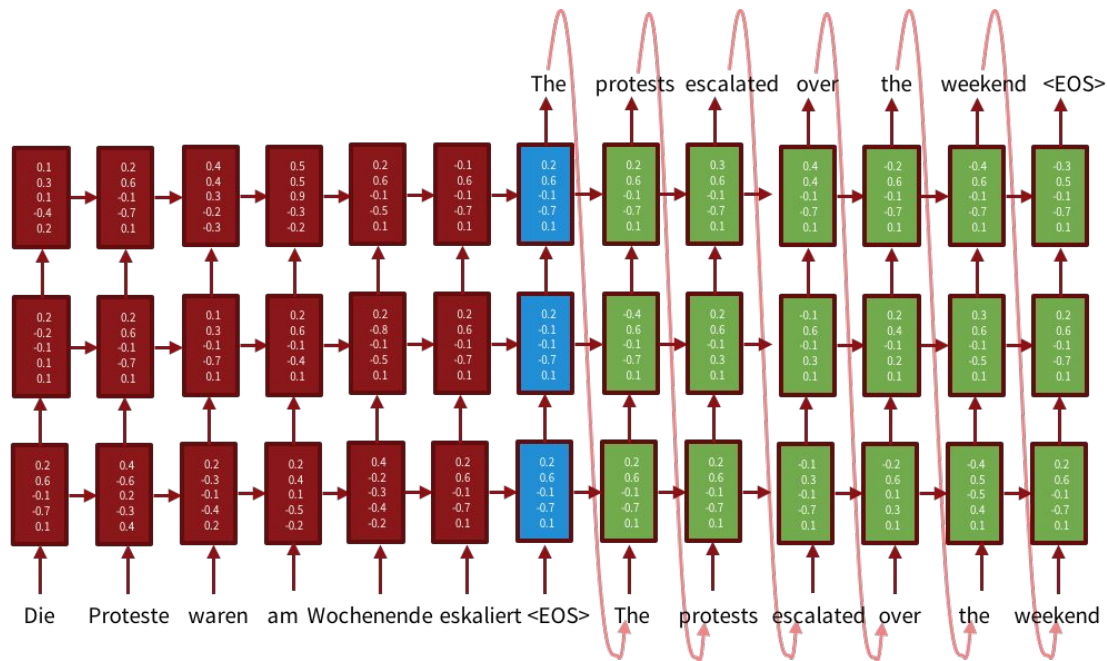
- **Cons:**

- Depending on **corpus** to compute probability, corpus creation make be costly
- **Specific errors** are hard to predict and fix
- Fluent result for short phrases, not for **large size sentence structures**

$$e^* = \arg \max_{e \in E} p(e) \times p(f|e)$$

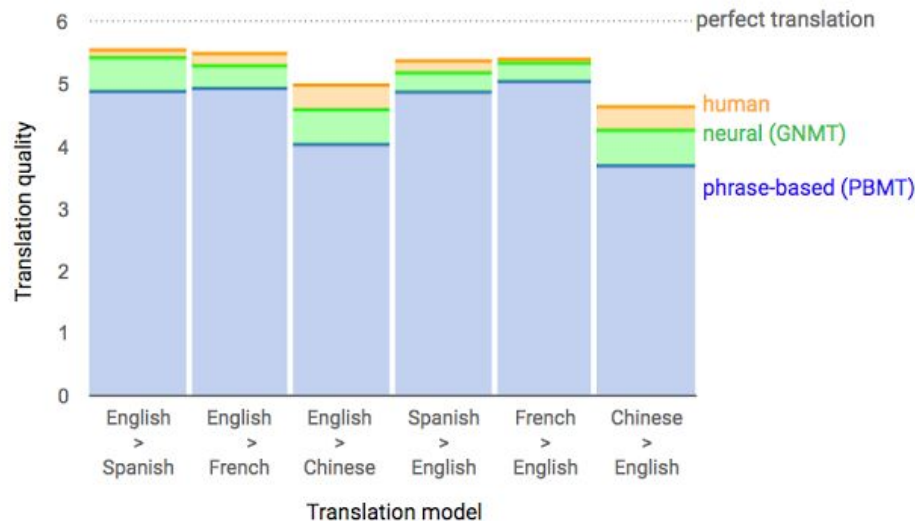


# Neural Machine Translation (NMT)



- **Training:** maximum likelihood estimation with backpropagation through time
- Vanishing gradient and gated recurrent units/long short-term memory units
- Conditional recurrent language modeling: **Encoder-Decoder**
- Decoding strategies

# Comparison of Different MT Approaches



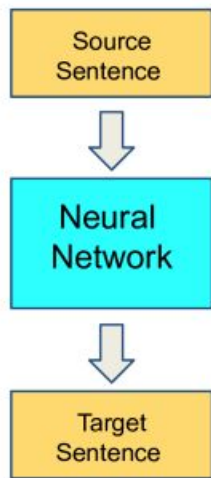
- Using human-rated side-by-side comparison as a metric, the *GNMT system produces translations that are vastly improved compared to the previous phrase-based production system*
- GNMT reduces translation errors by more than **55%-85%** on several major language pairs measured on sampled sentences from Wikipedia and news websites with the help of bilingual human raters

Input sentence:	Translation (PBMT):	Translation (GNMT):	Translation (human):
李克強此行將啟動中加總理年度對話機制，與加拿大總理杜魯多舉行兩國總理首次年度對話。	Li Keqiang premier added this line to start the annual dialogue mechanism with the Canadian Prime Minister Trudeau two prime ministers held its first annual session.	Li Keqiang will start the annual dialogue mechanism with Prime Minister Trudeau of Canada and hold the first annual dialogue between the two premiers.	Li Keqiang will initiate the annual dialogue mechanism between premiers of China and Canada during this visit, and hold the first annual dialogue with Premier Trudeau of Canada.

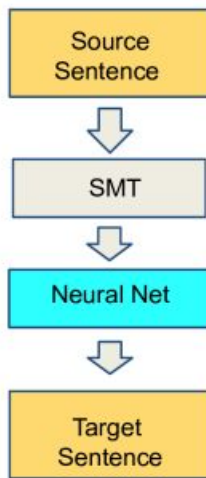
# Neural Machine Translation (NMT)



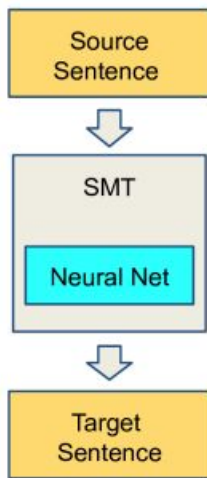
NMT so far achieves *better syntactic quality* than most SMT, but *poorer lexical quality*



Neural MT



(Schwenk et al. 2006)



(Devlin et al. 2014)

1. Devlin, Jacob, et al. *"Fast and Robust Neural Network Joint Models for Statistical Machine Translation."* ACL (1). 2014.
2. Liu, Shujie, et al. *"A recursive recurrent neural network for statistical machine translation."* (2014).
3. Wang, Xing, et al. *"Neural machine translation advised by statistical machine translation."* arXiv preprint arXiv:1610.05150 (2016).
4. Farajian, M. Amin, et al. *"Neural vs. Phrase-Based Machine Translation in a Multi-Domain Scenario."* EACL 2017 (2017): 280.



Thank you!

