

# Statistical Machine Translation

Hongning Wang

CS@UVa

# Machine translation



The image shows a screenshot of the Google Translate web interface. At the top, there are language selection buttons: English, Spanish, French, and French - detected (with a dropdown arrow). To the right of these is a swap button (two arrows). Below the language buttons is a text input box containing the following French text: "Le français est une langue indo-européenne de la famille des langues romanes. Le français s'est formé en France (variété de la « langue d'oïl ») et est aujourd'hui parlé sur tous les continents par environ 274 millions de personnes<sup>1</sup> dont 115 millions de locuteurs natifs<sup>3</sup>, auxquels s'ajoutent 72 millions de locuteurs partiels (évaluation Organisation internationale de la francophonie : 2010).". To the right of the text input box is a large blue arrow pointing downwards towards the output box. To the right of the arrow is the Google Translate logo. Below the text input box is a row of language selection buttons: Arabic, Chinese (Simplified), English (with a dropdown arrow), and a blue Translate button. Below the Translate button is a text output box containing the following English text: "French is an Indo-European language family of Romance languages. The French formed in France (variety of 'langue d'oïl') and is spoken today on every continent by about 274 million of which 115 million people<sup>1</sup> natifs<sup>3</sup> speakers, plus 72 million speakers partial (International Organization of the Francophonie evaluation: 2010).". At the bottom of the output box are icons for star, list, edit, and speaker.

English Spanish French French - detected

Le français est une langue indo-européenne de la famille des langues romanes. Le français s'est formé en France (variété de la « langue d'oïl ») et est aujourd'hui parlé sur tous les continents par environ 274 millions de personnes<sup>1</sup> dont 115 millions de locuteurs natifs<sup>3</sup>, auxquels s'ajoutent 72 millions de locuteurs partiels (évaluation Organisation internationale de la francophonie : 2010).

Google Translate

Arabic Chinese (Simplified) English

Translate

French is an Indo-European language family of Romance languages. The French formed in France (variety of "langue d'oïl") and is spoken today on every continent by about 274 million of which 115 million people<sup>1</sup> natifs<sup>3</sup> speakers, plus 72 million speakers partial (International Organization of the Francophonie evaluation: 2010).

# How do human translate languages?

- Is a bilingual dictionary sufficient?

John loves Mary.

Jean aime Marie.



John told Mary a story.

Jean a raconté une histoire à Marie.



John is ~~a~~ computer scientist.

Jean est informaticien.



John swam across the lake.

Jean a traversé le lac à la nage.



# Correspondences

**A bilingual dictionary is clearly insufficient!**

- One-to-one
  - John = Jean, aime = loves, Mary=Marie
- One-to-many/many-to-one
  - Mary = [à Marie]
  - [a computer scientist] = informaticien
- Many-to-many
  - [swam across \_\_\_] = [a traversé \_\_\_ à la nage]
- Reordering required
  - told Mary<sup>1</sup> [a story]<sup>2</sup> = a raconté [une histoire]<sup>2</sup> [à Marie]<sup>1</sup>

# Lexical divergences

- Different senses of homonymous words generally have different translations

English	- German
(river) bank	- Ufer
(financial) bank	- Bank

- Different senses of polysemous words may also have different translations

I **know** that he bought the book: Je **sais qu'**il a acheté le livre.

I **know** Peter: Je **connais** Peter.

I **know** math: Je **m'y connais en** maths.

# Syntactic divergences

- Word order
  - SVO (Sbj-Verb-Obj), SOV, VSO,...
  - fixed or free?
- Head-marking vs. dependent-marking
  - Dependent-marking (English): the man's house
  - Head-marking (Hungarian): the man house-his
- Pro-drop languages can omit pronouns
  - Italian (with inflection): I eat = mangio; he eats = mangia
  - Chinese (without inflection): I/he eat: chīfàn

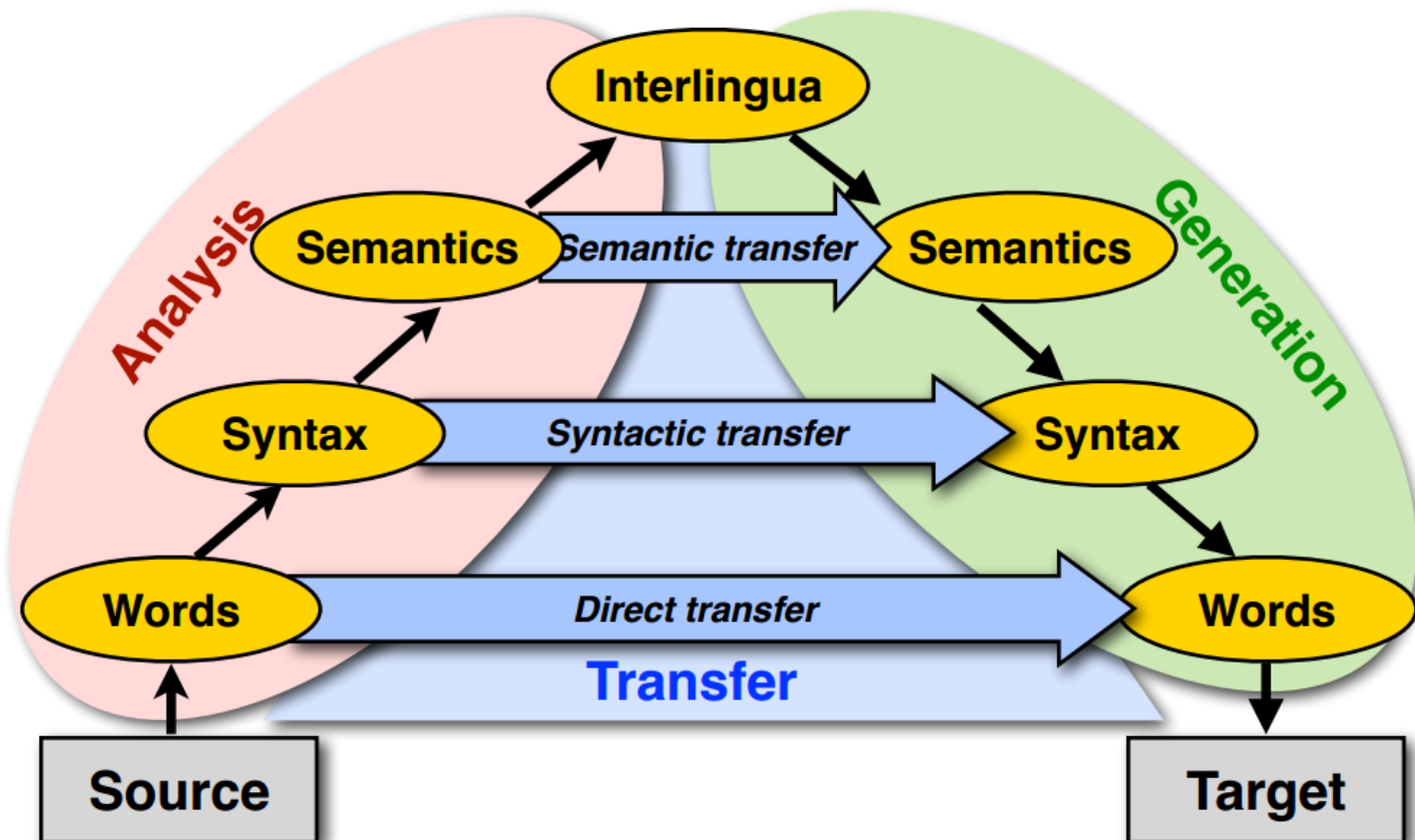
# Semantic differences

- Aspect:
  - English has a progressive aspect
    - ‘Peter swims’ vs. ‘Peter is swimming’
  - German can only express this with an adverb:
    - ‘Peter schwimmt’ vs. ‘Peter schwimmt gerade’

***Clearly, a bilingual dictionary is insufficient;  
and machine translation is difficult!***

# Machine translation approaches

- The Vauquois triangle





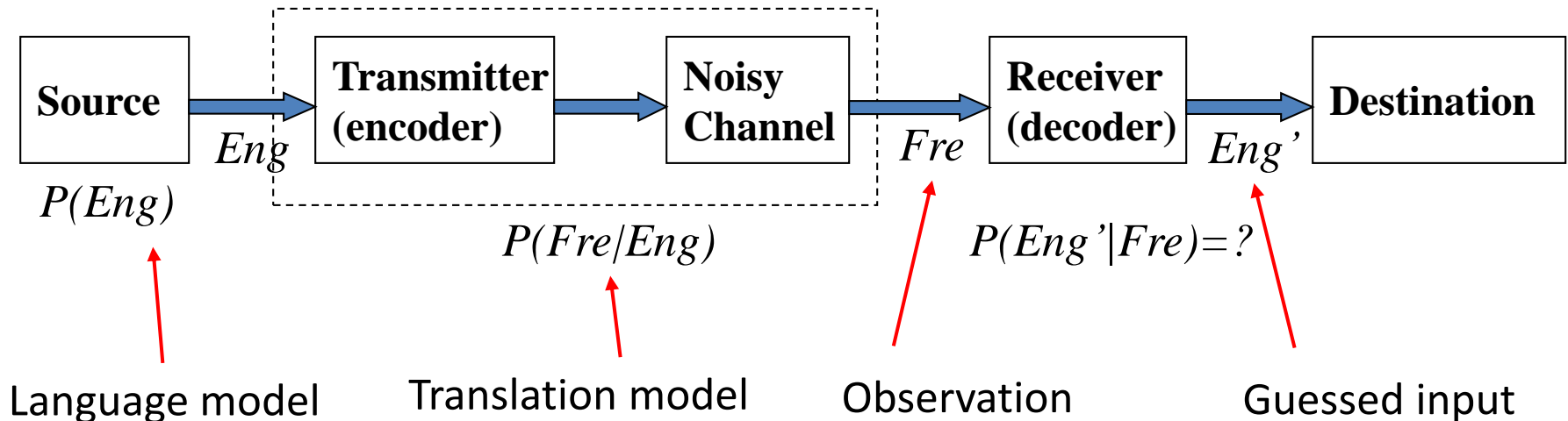
# Statistical machine translation

- Main stream of current machine translation paradigm
  - The idea was introduced by Warren Weaver in 1949
  - Re-introduced in 1993 by researchers at IBM's Thomas J. Watson Research Center
  - Now it is the most widely studied/used machine translation method

*1966: ALPAC report: human translation is far cheaper and better - kills MT for a long time*

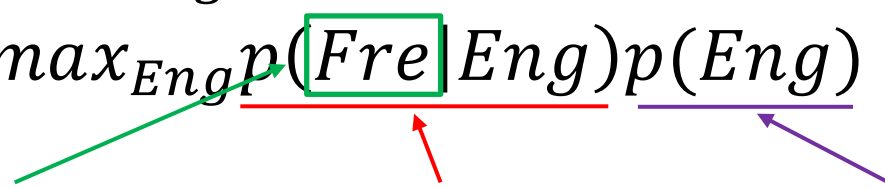
# Noisy-Channel framework [Shannon 48]

- Translating French to English
  - $Eng^* = \operatorname{argmax}_{Eng} p(Eng|Fre)$



# Translation with a noisy channel model

- Bayes rule

$$\begin{aligned} - Eng^* &= \operatorname{argmax}_{Eng} p(Eng|Fre) \\ &= \operatorname{argmax}_{Eng} p(\boxed{Fre} Eng) p(Eng) \end{aligned}$$


Observed (given)

Translation Model

Language Model

- Translation model  $p(Fre|Eng)$  should capture the **faithfulness** of the translation. It needs to be trained on *a parallel corpus*
- Language model  $p(Eng)$  should capture the **fluency** of the translation. It can be trained on *a very large monolingual corpus*

# Parallel Corpora

- The same text in two (or more) languages
  - High-quality manually crafted translations

## European Parliament Proceedings Parallel Corpus

- [parallel corpus Bulgarian-English](#), 41 MB, 01/2007-11/2011
- [parallel corpus Czech-English](#), 60 MB, 01/2007-11/2011
- [parallel corpus Danish-English](#), 179 MB, 04/1996-11/2011
- [parallel corpus German-English](#), 189 MB, 04/1996-11/2011
- [parallel corpus Greek-English](#), 145 MB, 04/1996-11/2011
- [parallel corpus Spanish-English](#), 187 MB, 04/1996-11/2011
- [parallel corpus Estonian-English](#), 57 MB, 01/2007-11/2011
- [parallel corpus Finnish-English](#), 179 MB, 01/1997-11/2011
- [parallel corpus French-English](#), 194 MB, 04/1996-11/2011
- [parallel corpus Hungarian-English](#), 59 MB, 01/2007-11/2011
- [parallel corpus Italian-English](#), 188 MB, 04/1996-11/2011
- [parallel corpus Lithuanian-English](#), 57 MB, 01/2007-11/2011
- [parallel corpus Latvian-English](#), 57 MB, 01/2007-11/2011
- [parallel corpus Dutch-English](#), 190 MB, 04/1996-11/2011
- [parallel corpus Polish-English](#), 59 MB, 01/2007-11/2011
- [parallel corpus Portuguese-English](#), 189 MB, 04/1996-11/2011
- [parallel corpus Romanian-English](#), 37 MB, 01/2007-11/2011
- [parallel corpus Slovak-English](#), 59 MB, 01/2007-11/2011
- [parallel corpus Slovene-English](#), 54 MB, 01/2007-11/2011
- [parallel corpus Swedish-English](#), 171 MB, 01/1997-11/2011

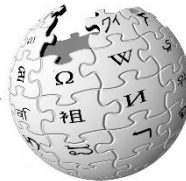
# Parallel Corpora

- The same text in two (or more) languages
  - High-quality manually crafted translations

## Alan Turing

From Wikipedia, the free encyclopedia


WIKIPEDIA  
The Free Encyclopedia



*"Turing" redirects here. For other uses, see [Turing \(disambiguation\)](#)*

**Alan Mathison Turing**, OBE, FRS (/ˈtʃʊərɪn/ *TEWR-ing*; 23 June 1912 – 7 June 1954) was a British pioneering [computer scientist](#), mathematician, [logician](#), [cryptanalyst](#), philosopher, mathematical biologist, and marathon and ultra distance runner. He was highly influential in the development of [computer science](#), providing a formalisation of the concepts of "[algorithm](#)" and "[computation](#)" with the [Turing machine](#), which can be considered a model of a general purpose computer.<sup>[3][4][5]</sup> Turing is widely considered to be the father of theoretical computer science and [artificial intelligence](#).<sup>[6]</sup>

## Alan Turing

 *Pour les articles homonymes, voir [Turing](#).*

**Alan Mathison Turing**, OBE, FRS (23 juin 1912 - 7 juin 1954), est un [mathématicien](#), [cryptologue](#) et [informaticien britannique](#).

Il est l'auteur, en 1936, d'un article de logique mathématique<sup>1</sup> qui est devenu plus tard un texte fondateur de la science [informatique](#). Pour résoudre le problème fondamental de la décidabilité en arithmétique, il y présente une [expérience de pensée](#) que l'on nommera ensuite [machine de Turing](#) et des concepts de [programmation](#) et de [programme](#)<sup>2,3</sup>, qui prendront tout leur sens avec la diffusion des [ordinateurs](#), dans la seconde moitié du xx<sup>e</sup> siècle. Avec

# Parallel Corpora

- The same text in two (or more) languages
  - High-quality manually crafted translations



## Cosmo

Où sont les filles, les femmes au tempérament de guerrière  
Oui qui savent comment faire la fête, qu'elles soient mère ou célibataires  
Où sont les hommes, les gangstes,  
Les pauvres ou les millionnaires  
Les bobos, les mecs en survet'  
Les intellos, les mecs en fumette,  
Où sont les quartiers, les blocs,  
Les HLM mis de côtés,  
Les résidences les quartiers huppés,  
Les 205, les AUDI TT  
Où sont les blacks, les blancs, les jaunes, les verts, les rouges et les gris  
Loin des amalgames politiques  
Bienvenue en Cosmopolitanie

## Cosmo

Where are the girls, the women with a warrior temperament  
Yes who know how to party, no matter if they're mothers or singles  
Where are the men, the gangsters,  
The poor or the millionaires  
The bobos, the guys in tracksuit,  
The nerds, the guys smoking joints,  
Where are the districts, the blocks,  
The social housing put aside,  
The residences the posh districts,  
The 205\*, the AUDI TT\*  
Where are the Blacks, the Whites, the Yellows, the Greens, the Reds and the Greys  
Far from political amalgamation  
Welcome in Cosmopolitany

# Translation Model $p(Fre|Eng)$

- Specifying translation probabilities

English	French	Frequency
green witch	grüne Hexe	...
at home	zu Hause	10534
at home	daheim	9890
is	ist	598012
this week	diese Woche	...

- This probability needs word-alignment to estimate

# Language model $p(\text{Eng})$

- Specifying the likelihood of observing a sentence in the target language
  - N-gram language model
    - Relax the language complexity
    - Occurrence of current word only depends on previous N-1 words:  $p(w_1 \dots w_n) = \prod_i p(w_i | w_{i-1}, \dots, w_{i-N+1})$



# Language model $p(\text{Eng})$

- Specifying the likelihood of observing a sentence in the target language
  - Google (2007) uses 5-grams to 7-grams, which result in huge models, but the effect on translation quality levels off quickly

## Size of models

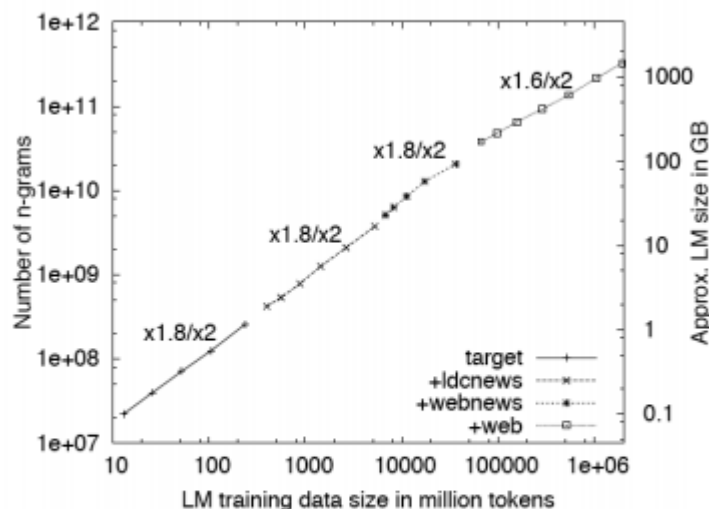


Figure 3: Number of  $n$ -grams (sum of unigrams to  $n$ -grams) for varying amounts of training data.

## Effect on translation quality

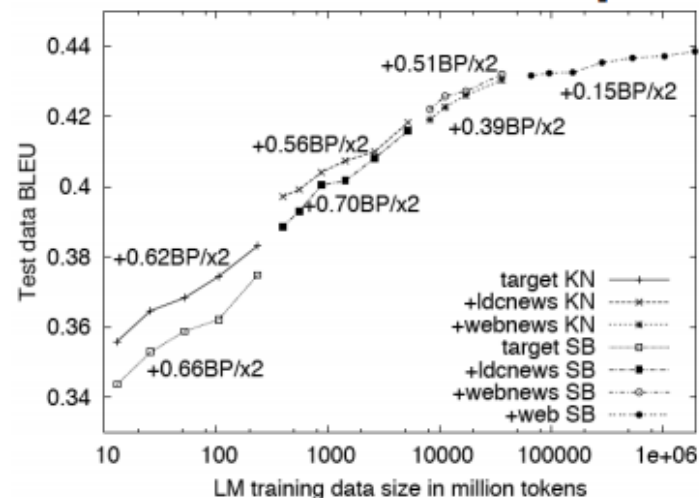
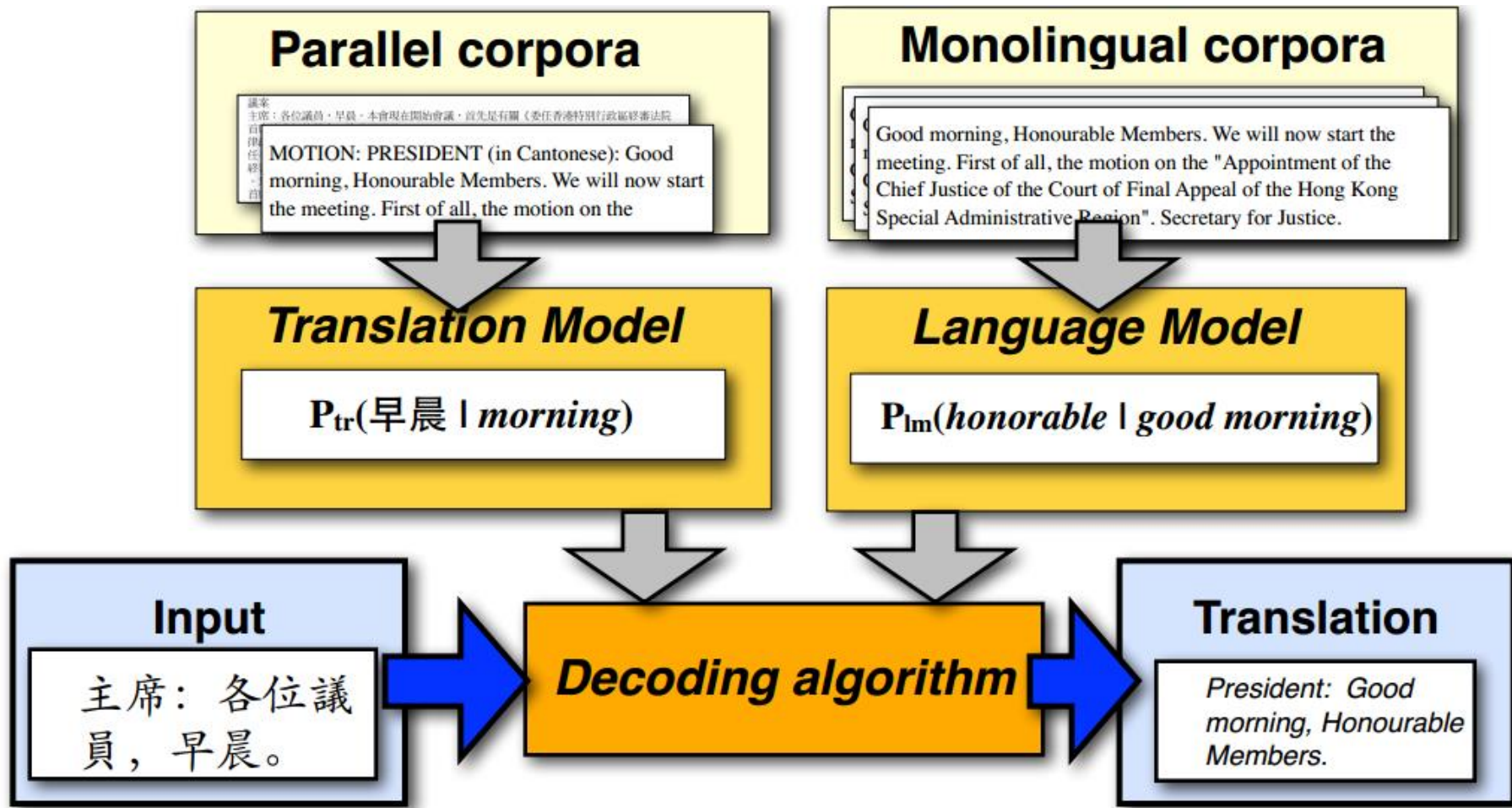


Figure 5: BLEU scores for varying amounts of data using Kneser-Ney (KN) and Stupid Backoff (SB).

# Statistical machine translation



# IBM translation models

- A generative model based on noisy channel framework
  - Generate the translation sentence  $e$  with regard to the given sentence  $f$  by a stochastic process
    1. Generate the length of  $f$
    2. Generate the **alignment** of  $e$  to the target sentence  $f$
    3. Generate the words of  $f$
  - $Eng^* = \operatorname{argmax}_{Eng} p(Fre|Eng)p(Eng)$

# Word alignment

- One to many

John told Mary a story.  
Jean a raconté une histoire à Marie.

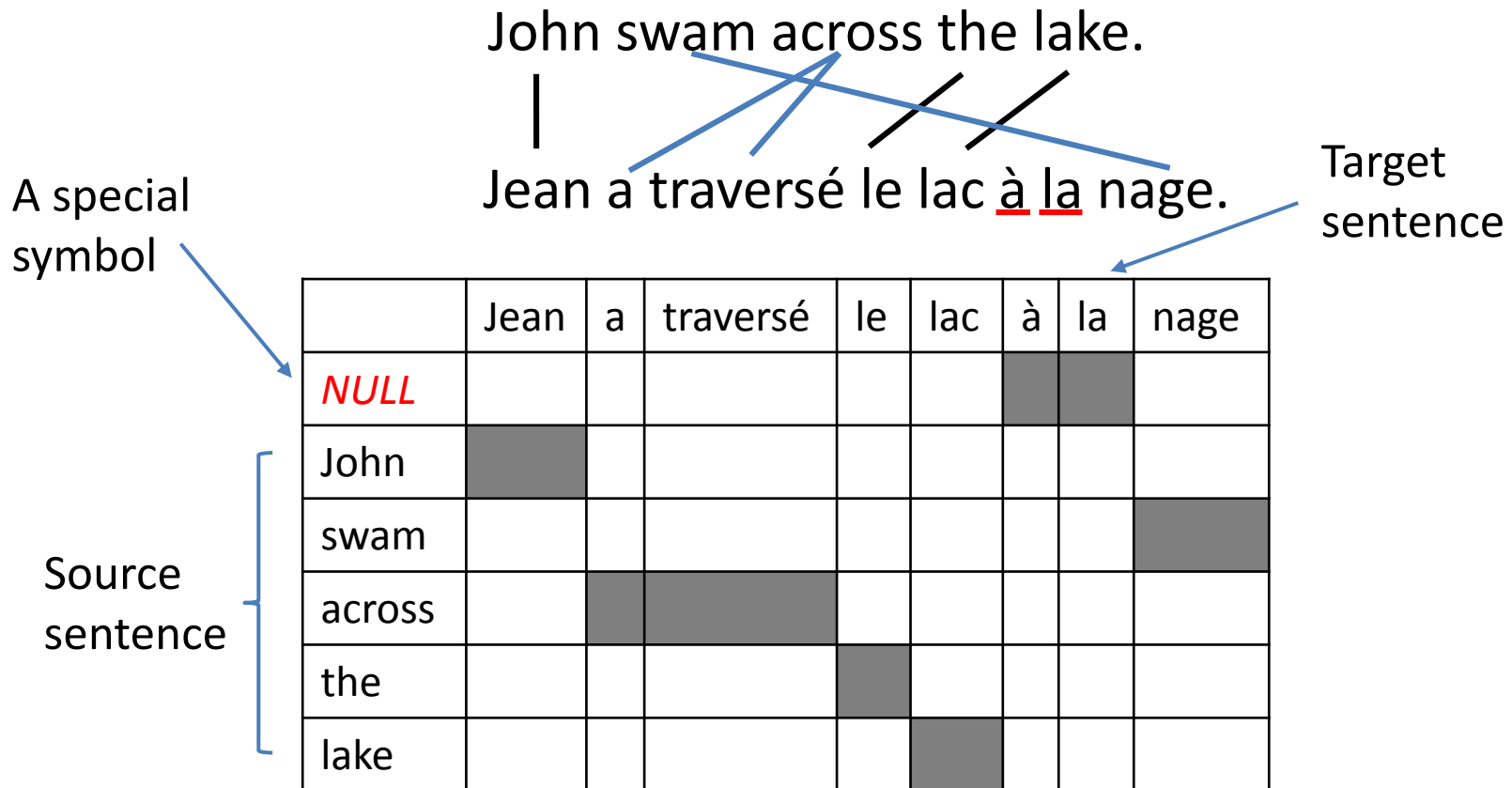
Target sentence

Source sentence

	Jean	a	raconté	une	histoire	à	Marie
John							
told							
Mary							
a							
story							

# Word alignment

- Many to one and missing word



# Representing word alignments

- Alignment table

		1	2	3	4	5	6	7	8
		Jean	a	traversé	le	lac	à	la	nage
0	<i>NULL</i>								
1	John								
2	swam								
3	across								
4	the								
5	lake								



Target Position	1	2	3	4	5	6	7	8
Source Position	1	3	3	4	5	0	0	2

# IBM translation models

- Translation model with word alignment

$$- p(Fre|Eng) = \sum_{a \in A(Eng, Fre)} p(Fre, a|Eng)$$

*marginalize over all possible alignments  $a$*

- Generate the words of  $f$  with respect to alignment  $a$

$$p(\mathbf{f}, \mathbf{a} | \mathbf{e}) = p(m | \mathbf{e}) \prod_{j=1}^m p(a_j | a_{1..j-1}, f_{1..j-1}, m, \mathbf{e}) p(f_j | a_{1..j}, f_{1..j-1}, m, \mathbf{e})$$

Length of target sentence  $f$

Word alignment  $a_j$


Translation of  $f_j$

# IBM translation models

- Sequence of 5 translation models
  - Different assumptions and realization of the components in the translation models, i.e., length model, alignment model and translation model
  - Model 1 is the simplest and becomes the basis of follow-up IBM translation models




# Parameters in Model 1

- Length probability  $p(m|\mathbf{e})$ 
  - Probability of generating a source sentence of length  $m$  given a target sentence  $\mathbf{e}$ 
    - Assumed to be constant -  $p(m|\mathbf{e}) = \epsilon$
- Alignment probability  $p(a|\mathbf{e})$ 
  - Probability of source position  $i$  is aligned to target position  $j$ 
    - Assumed to be uniform -  $p(a|\mathbf{e}) = \frac{1}{n}$   *length of source sentence*

# Parameters in Model 1

- Translation probability  $p(f|a, e)$ 
  - Probability of English word  $e_i$  is translated to French word  $f_j$  -  $p(f_j|e_{a_j})$
- After the simplification, Model 1 becomes

$$\begin{aligned} p(\mathbf{f}, \mathbf{a}|\mathbf{e}) &= p(m|\mathbf{e}) \prod_{j=1}^m p(a_j|a_{1..j-1}, f_{1..j-1}, m, \mathbf{e}) p(f_j|a_{1..j}, f_{1..j-1}, m, \mathbf{e}) \\ &= \frac{\epsilon}{(n+1)^m} \prod_{j=1}^m p(f_j|e_{a_j}) \end{aligned}$$


We add a NULL word in the source sentence

# Generative process in Model 1

For a particular English sentence  $e = e_1..e_n$  of length  $n$

0	1	2	3	4	5
NULL	John	swam	across	the	lake

1. Choose a length  $m$  for the target sentence (e.g  $m = 8$ )

1	2	3	4	5	6	7	8
Jean	a	traversé	le	lac	à	la	nage

2. Choose an alignment  $a = a_1 ... a_m$  for the source sentence

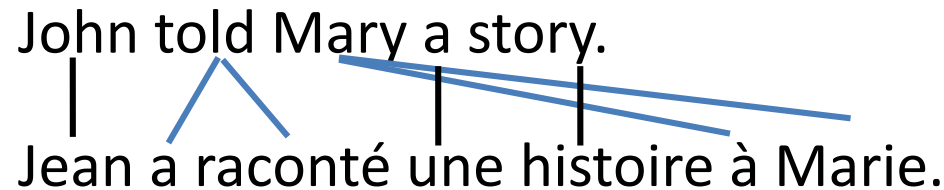
Target Position	1	2	3	4	5	6	7	8
Source Position	1	3	3	4	5	0	0	2

3. Translate each source word  $e_{a_j}$  into the target language

English	John	across	across	the	lake	NULL	NULL	swam
Alignment	1	3	3	4	5	0	0	2
Translation	Jean	a	traversé	le	lac	à	la	nage

# Estimation of translation probability

- If we have ground-truth word-alignments in the parallel corpus, maximum likelihood estimator is sufficient



$$- p(f|e) = \frac{c(e \rightarrow f)}{\sum_w c(e \rightarrow w)}$$

# Estimation of translation probability

- If we do not have ground-truth word-alignments, appeal to Expectation Maximization algorithm
  - Intuitively, guess the alignment based on the current translation probability first; and then update the translation probability
  - EM algorithm will be carefully discussed in our later lecture of “Text Clustering”

# Other translation models

- IBM models 2-5 are more complex
  - Word order and string position of the aligned words
  - Phrase translation in the source and target languages