Statistical Machine Translation

Hongning Wang
CS@UVa

Machine translation

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^1 [a story]^2 = a raconté [une histoire]^2 [à Marie]^1

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

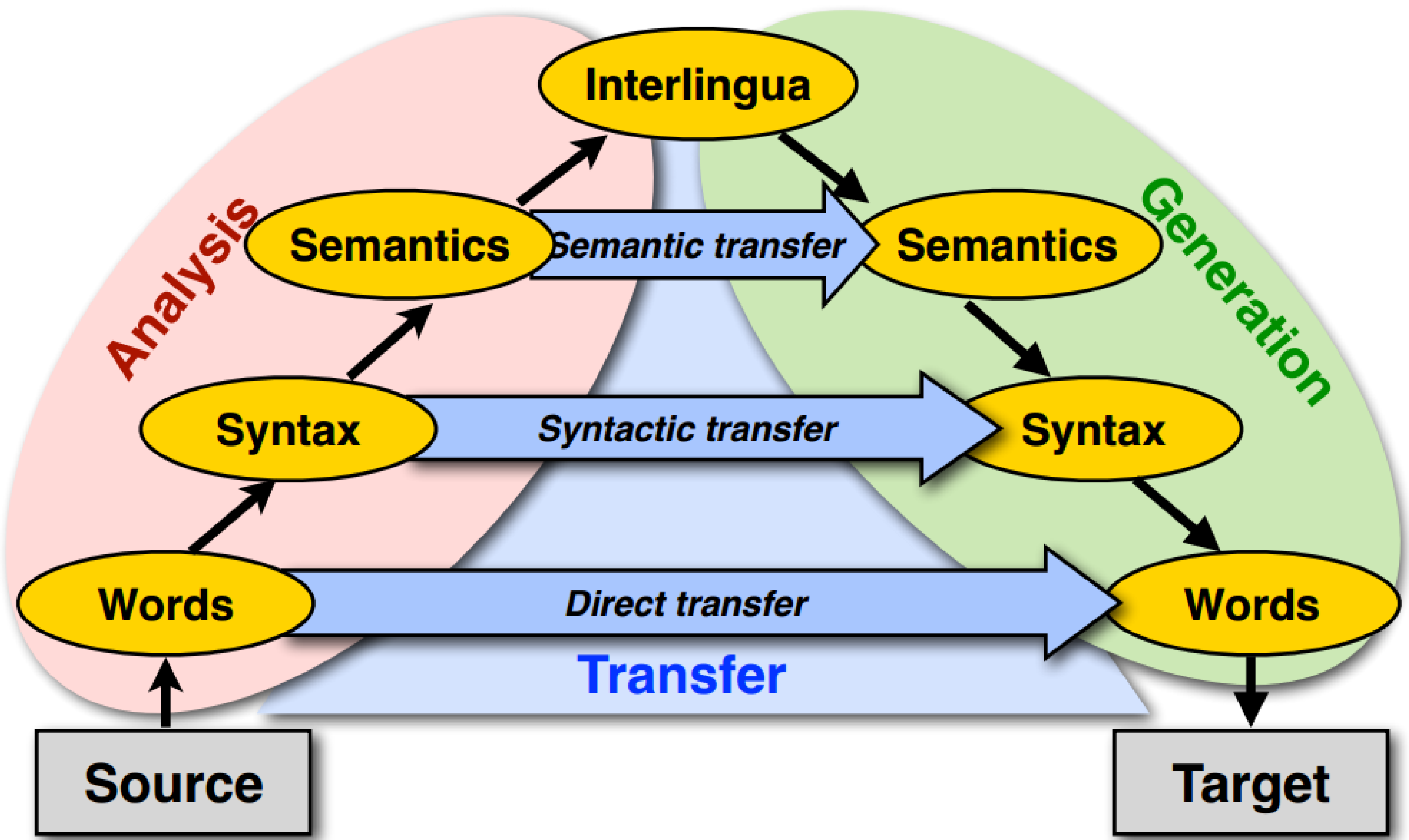
Semantic divergences

- 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



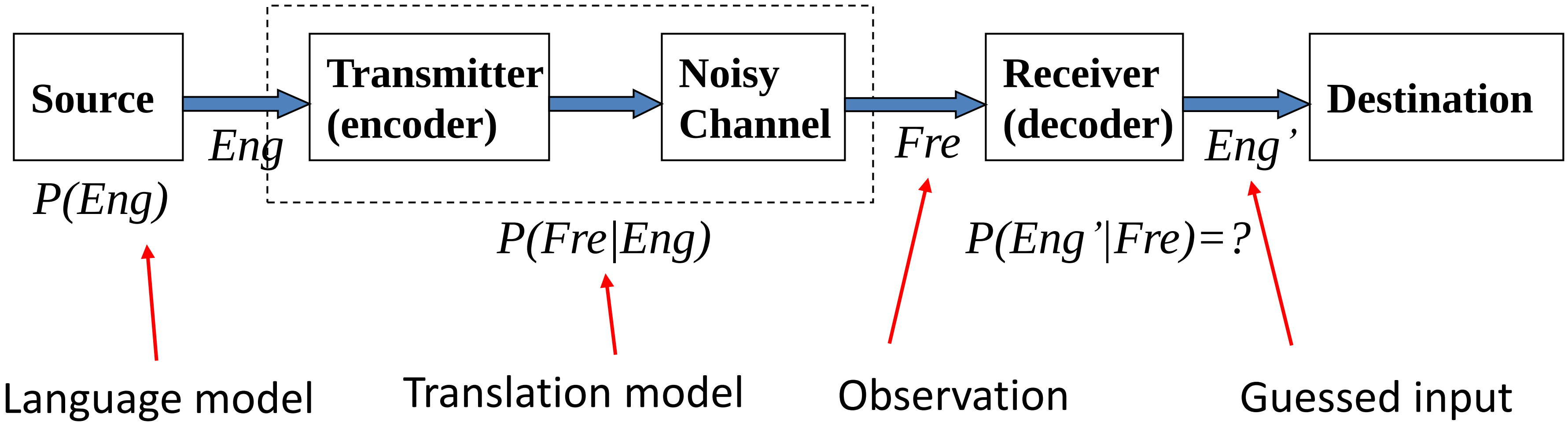
Statistical machine translation

- 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
  - $Eng^* = argmax_{Eng} p(Eng|Fre)$



Translation with a noisy channel model

- Bayes rule
  - $Eng^* = argmax_{Eng} p(Eng|Fre)$
    $= argmax_{Eng} \underbrace{p(\boxed{Fre}|Eng)}_{\text{Translation Model}} \underbrace{p(Eng)}_{\text{Language Model}}$

    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

WIKIPEDIA
The Free Encyclopedia

Alan Turing

From Wikipedia, the free encyclopedia

"Turing" redirects here. For other uses, see Turing (disambiguation

Alan Mathison Turing, OBE, FRS (/ˈtjʊərɪŋ/ *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.[3][4][5] Turing is widely considered to be the father of theoretical computer science and artificial intelligence.[6]

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[1] 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[2, 3], qui prendront tout leur sens avec la diffusion des ordinateurs, dans la seconde moitié du XXe siècle. Avec

Parallel corpora

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

LYRICS TRANSLATE

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	zuhause	10534
at home	daheim	9890
is	ist	598012
this week	diese Woche	…

  – This probability needs word-alignment to estimate

Language model p(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(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

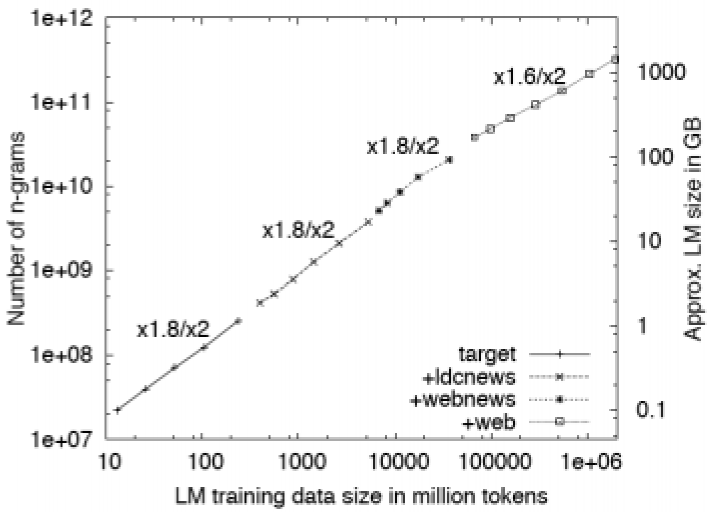


Figure 3: Number of n-grams (sum of unigrams to 5-grams) for varying amounts of training data.

Effect on translation quality

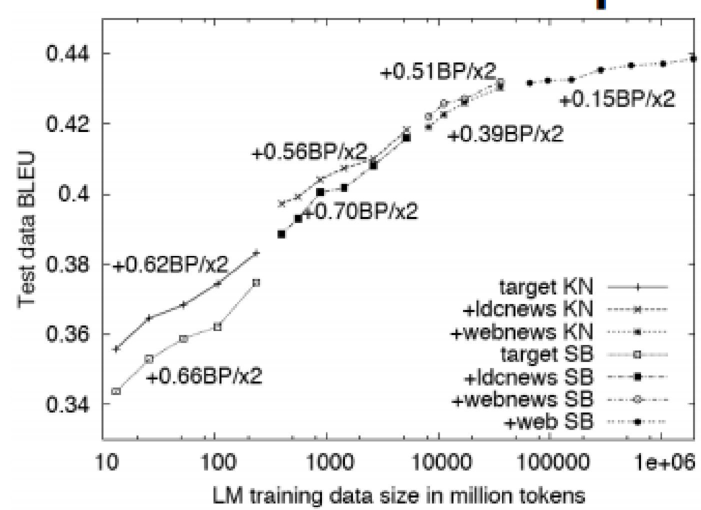
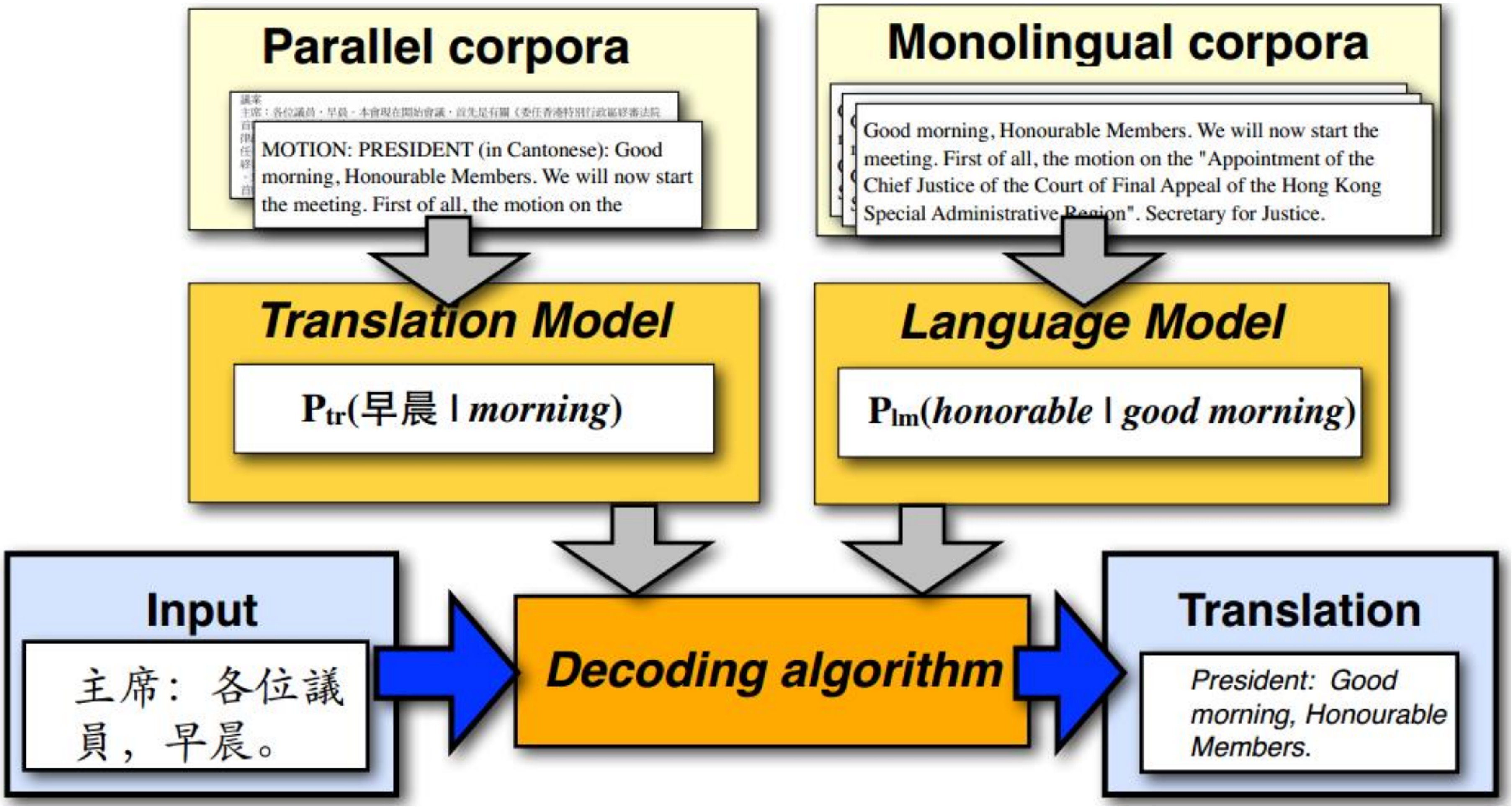


Figure 5: BLEU scores for varying amounts of data using Kneser-Ney (KN) and Stupid Backoff (SB).

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***
  - $Eng^* = argmax_{Eng} p(Fre|Eng)p(Eng)$

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

John swam across the lake.

Jean a traversé le lac <u>à</u> <u>la</u> nage.

A special symbol

Target sentence

Source sentence

	Jean	a	traversé	le	lac	à	la	nage
NULL								
John								
swam								
across								
the								
lake								

Representing word alignments

- Alignment table

		1	2	3	4	5	6	7	8
		Jean	a	traversé	le	lac	à	la	nage
0	*NULL*								
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 $\boldsymbol{f}$ with respect to alignment $\boldsymbol{a}$

$$p(\boldsymbol{f}, \boldsymbol{a}|\boldsymbol{e}) = \underbrace{\boxed{p(m|\boldsymbol{e})}}_{\text{Length of target sentence } \boldsymbol{f}} \prod_{j=1}^{m} \underbrace{\boxed{p\left(a_j \middle| a_{1..j-1}, f_{1,..j-1}, m, \boldsymbol{e}\right)}}_{\text{Word alignment } a_j} \underbrace{\boxed{p(f_j|a_{1..j}, f_{1,..j-1}, m, \boldsymbol{e})}}_{\text{Translation of } f_j}$$

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

Parameters in Model 1

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

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

$$= \frac{\epsilon}{(n+1)^m}\prod_{j=1}^{m} p(f_j|e_{a_j})$$

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
?	?	?	?	?	?	?	?

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

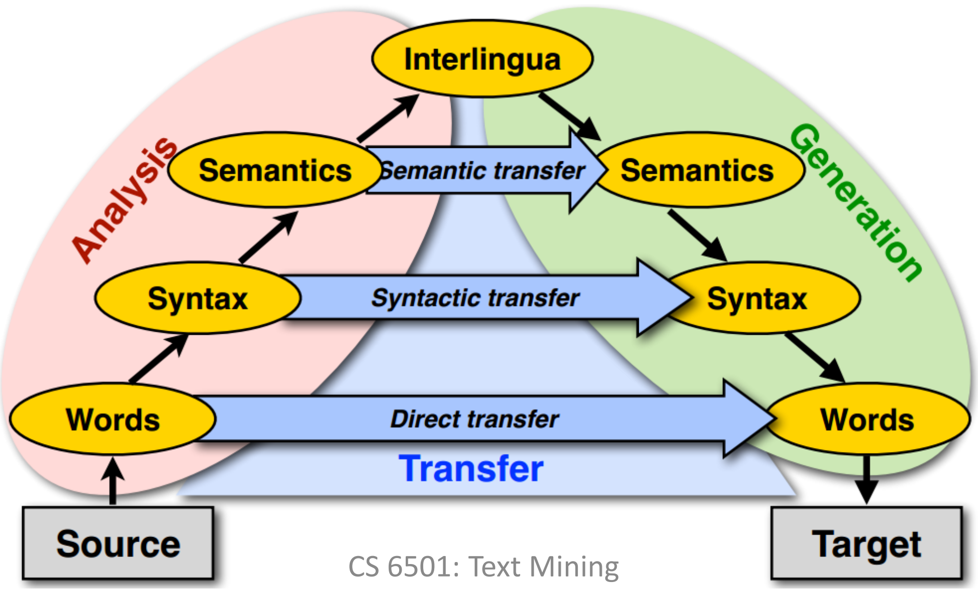
Target Position	1	2	3	4	5	6	7	8
Source Position	1	3	3	4	5	0	0	2

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

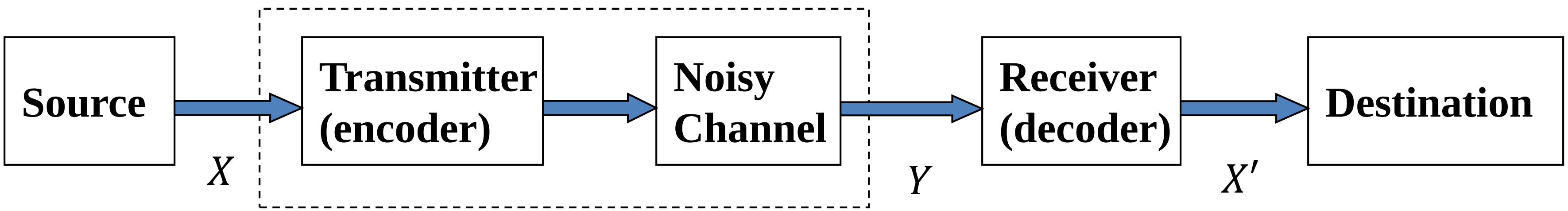
Recap: machine translation

- Challenges
  - Lexical, syntactic, semantic divergences
- Vauquois triangle
  - Ideal procedure for translation

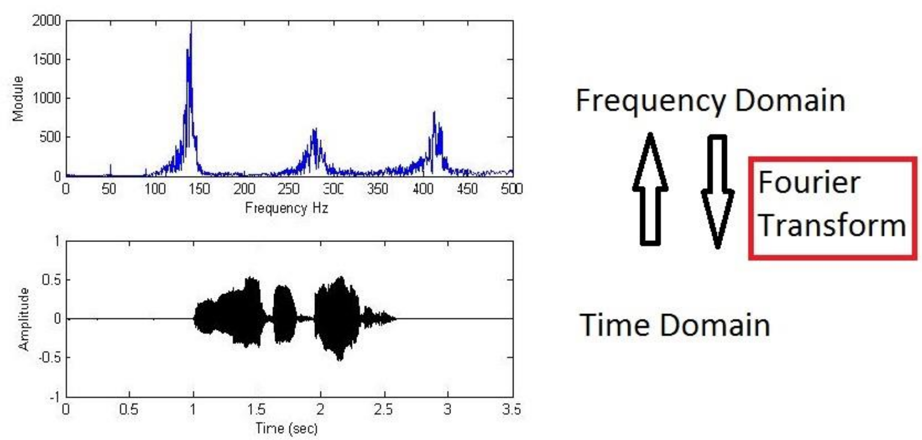


Recap: statistical machine translation

- Source-channel framework in signal processing

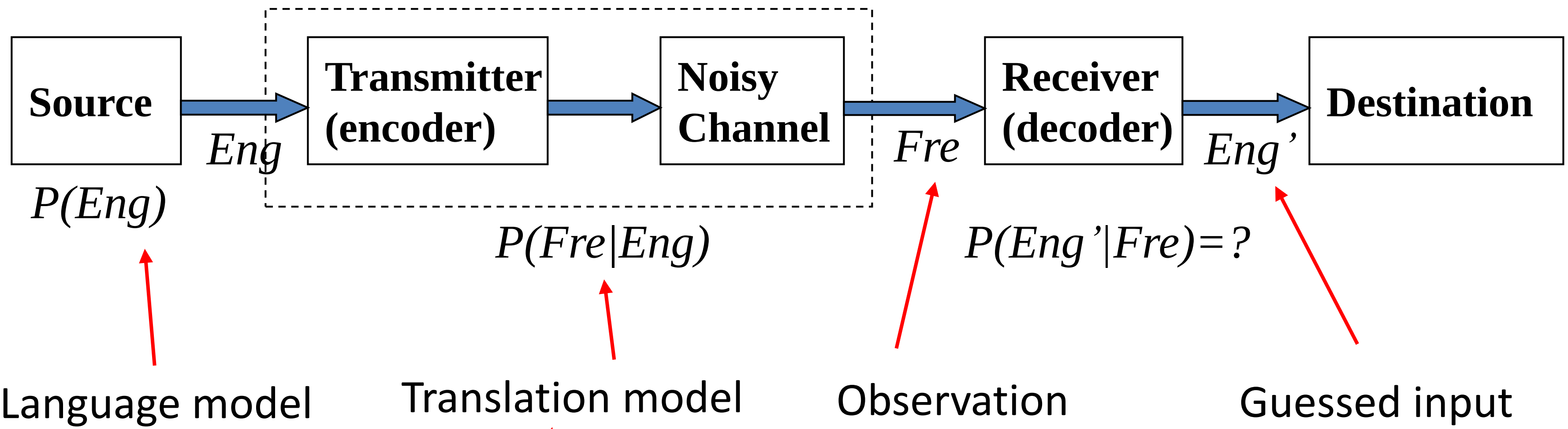


$Y = f(X)$ $\quad$ $X' = f^{-1}(Y)$



Recap: statistical machine translation

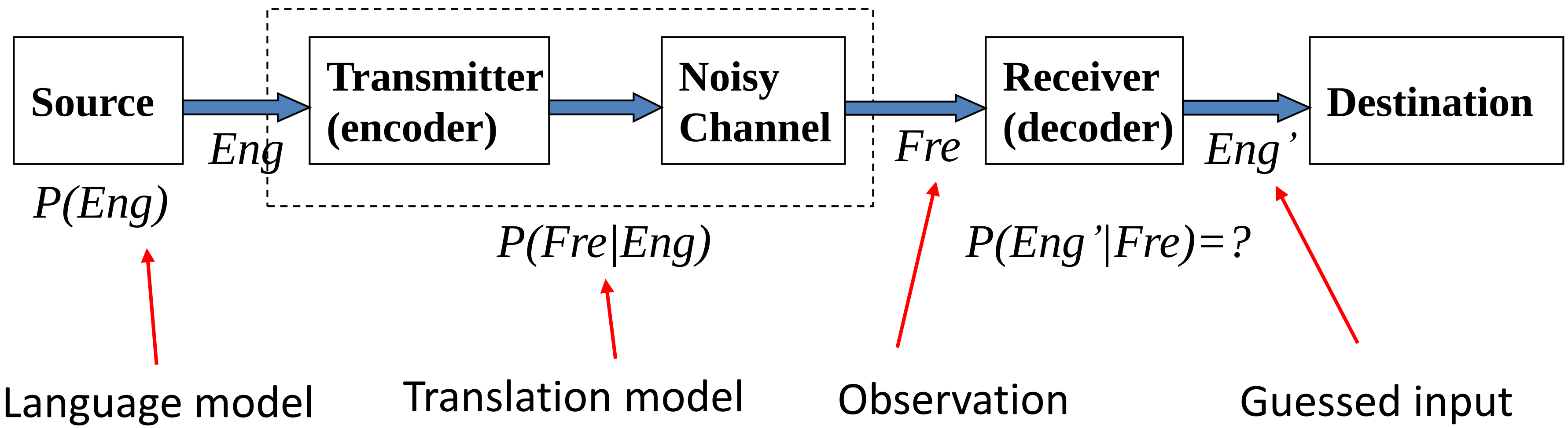
- Source-channel framework for translation
  - $Eng^* = argmax_{Eng} p(Eng|Fre)$



1. **Naturally decompose into two sub-problems;**
2. **Independence assumptions can be introduced.**

Recap: statistical machine translation

- Source-channel framework for translation
  - $Eng^* = argmax_{Eng} p(Eng|Fre)$

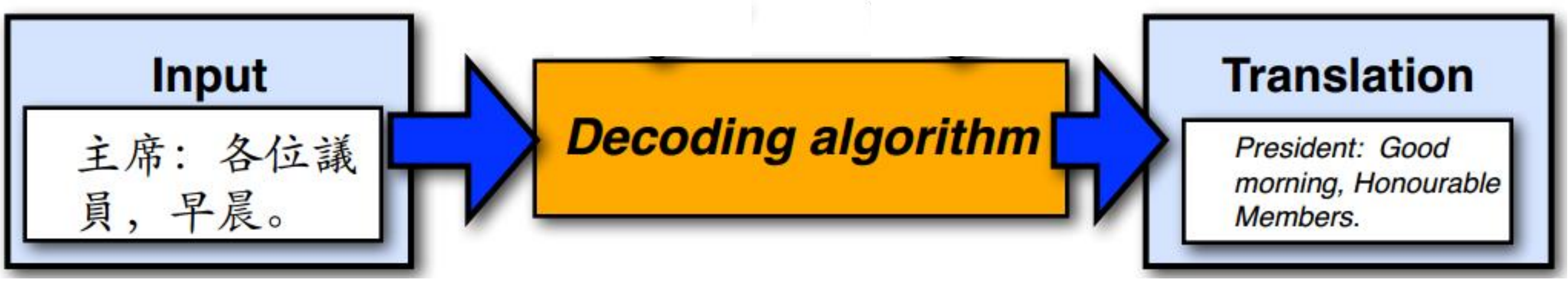


How to define inverse operation in a probabilistic model?

$f^{-1}(f(X)) = X$

$Eng^* = argmax_{Eng} p(Fre|Eng)p(Eng)$

Recap: statistical machine translation



Generative process in Model 1

Source

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

Transmitter

Order of action

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

1	2	3	4	5	6	7	8
?	?	?	?	?	?	?	?

2. Choose an alignment $a = a_1 \ldots a_m$ 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
Encoded	Jean	a	traversé	le	lac	à	la	nage

Decoding process in Model 1

$p(\boldsymbol{e}|\boldsymbol{f}) = 1e^{-55}$

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

$p(\boldsymbol{e})$

0	1	2	3	4	5
NULL	**John**	**flies**	**across**	**the**	**river**

Search through all English sentences

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

$p(m|\boldsymbol{e}) = \epsilon$

1	2	3	4	5	6	7	8
?	?	?	?	?	?	?	?

Search through all possible alignments

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

$p(a|\boldsymbol{e}) = \frac{1}{n}$

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

Order of action

$\prod_{j=1}^{m} p(f_j|e_{a_j})$

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

English	**John**	**flies**	**the**	**river**	**river**	**flies**	**NULL**	**across**
Alignment	**1**	**2**	**4**	**5**	**5**	**2**	**0**	**3**
Encoded	Jean	a	traversé	le	lac	à	la	nage

Receiver

Decoding process in Model 1

$p(\boldsymbol{e}|\boldsymbol{f}) = 1e^{-15}$

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

$p(\boldsymbol{e})$

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

Search through all English sentences

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

$p(m|\boldsymbol{e}) = \epsilon$

1	2	3	4	5	6	7	8
?	?	?	?	?	?	?	?

Search through all possible alignments

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

$p(a|\boldsymbol{e}) = \frac{1}{n}$

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

Order of action

$\prod_{j=1}^{m} p(f_j|e_{a_j})$

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**
Encoded	Jean	a	traversé	le	lac	à	la	nage

Receiver

Decoding process in Model 1

- Search space is huge
  – Presumably all “sentences” in English
    - English sentence length is unknown
    - All permutation of words in the vocabulary
  – Heuristics to reduce search space
    - Trade-off between translation accuracy and efficiency

Estimation of translation probability

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

John told Mary a story.

Jean a raconté une histoire à Marie.

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

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
  - Phase-based translation in the source and target languages
    - Incorporate syntax or quasi-syntactic structures
    - Greatly reduce search space

What you should know

- Challenges in machine translation
  – Lexicon/syntactic/semantic divergences
- Statistical machine translation
  – Source-channel framework for statistical machine translation
    - Generative process
  – IBM model 1
    - Idea of word alignment

Today's reading

- Speech and Language Processing
  - Chapter 25: Machine Translation