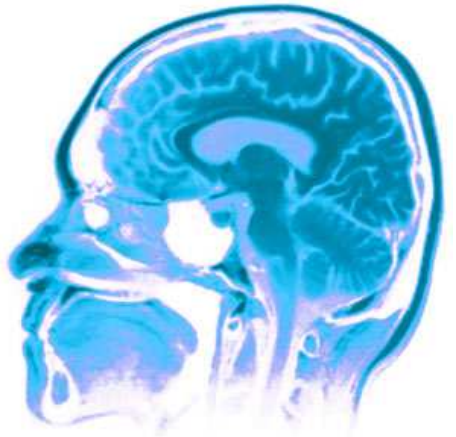
Machine Translation



Shafiq Joty

Scientist @ QCRI

NTU-SCSE

Oct-19, 2016

About me

Shafiq Joty

- Scientist at Qatar Computing Research Institute
- PhD from University of British Columbia, Canada
- 10 years in Natural Language Processing
  - Discourse
  - NLP Applications (MT, Summarization, QA)
- 3 Years in Machine Translation and Its Evaluation
  - Arabic – English
  - German - English

Machine Translation (MT)

- Classic test of language understanding
  Language *analysis* and *generation*

- German <=> English

Automatische Textverarbeitung gefällt mir.

I like natural language processing.

Example from Cheung (2015)

Machine Translation (MT)

- Chinese <=> English

因为一项2011年赤字削减协议，如果无法与共和党达成折中方案，奥巴马总统可能在年底面临联邦预算自动减少1000多亿美元的局面。在外交政策辩论中，奥巴马说，他的军事预算不会“减少”而将“维持”。

A 2011 deficit reduction agreement, if a compromise can not be reached with the Republican Party, President Obama may face at the end of the federal budget situation automatically reduced by more than 1000 billion dollars. In the foreign policy debate, Obama said, his military budget will not "reduce" and "maintain".

Example from Koehn (2016)

The Need for Machine Translation

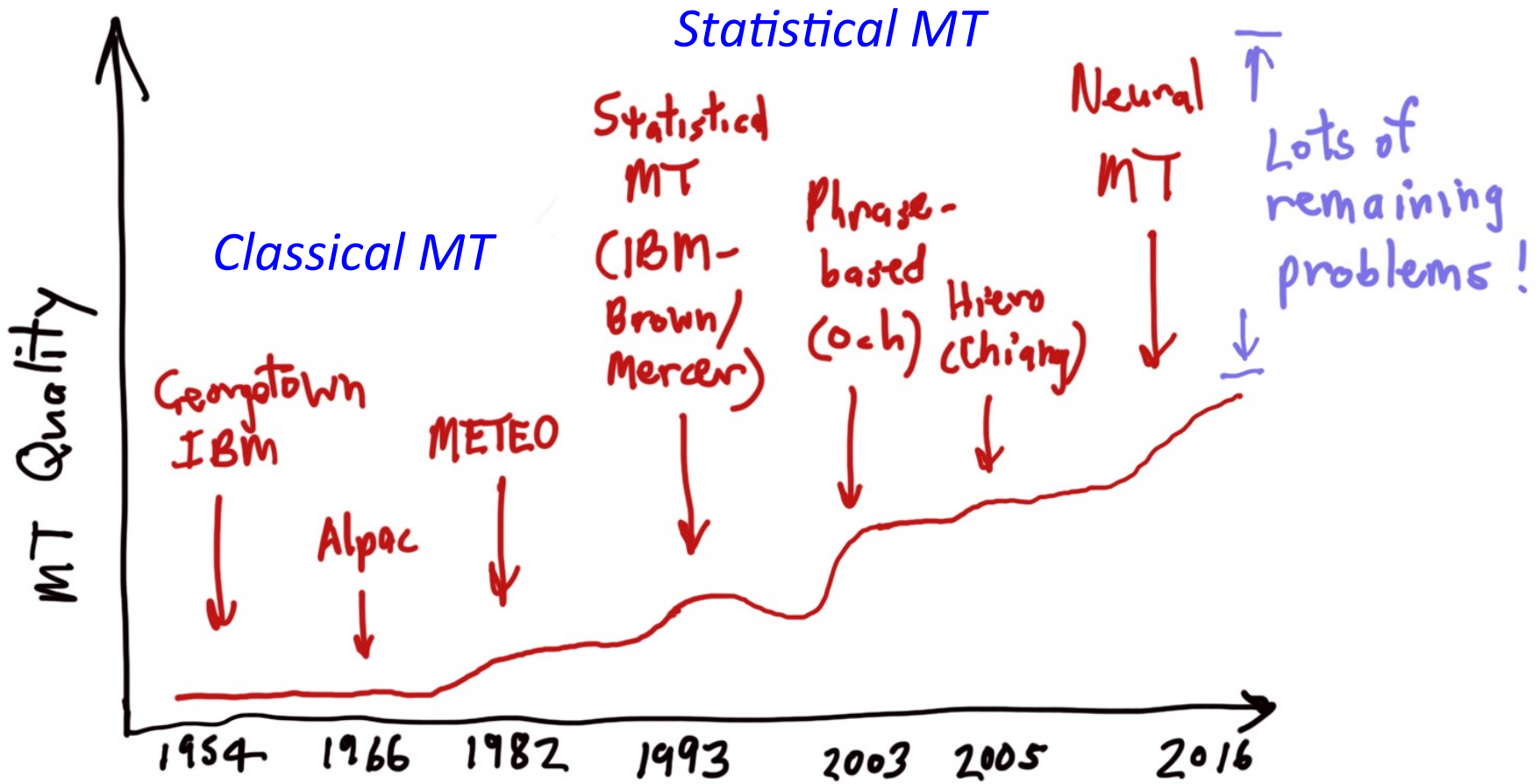
- Big MT needs: for humanity and for commerce
  - Translation is a US$40 billion a year industry
  - Huge in Europe, growing in Asia
  - Large social/government/military as well as commercial needs

- Huge commercial use
  - Google translates over 100 billion words a day
  - eBay uses MT to enable cross-border trade
  - Facebook has just rolled out new homegrown MT

    "When we turned [MT] off for some people, they went nuts!"

Source: Luong, Cho, Manning (ACL tutorial 2016)

Progress in MT



Source: Luong, Cho, Manning (ACL tutorial 2016)

Lecture Outline

- Challenges in MT
- Classical rule-based MT
- Statistical MT: a noisy channel model
- Statistical MT: log-linear model
- Neural MT

Why is MT Difficult?

Automatische Textverarbeitung gefällt mir.

I like natural language processing.

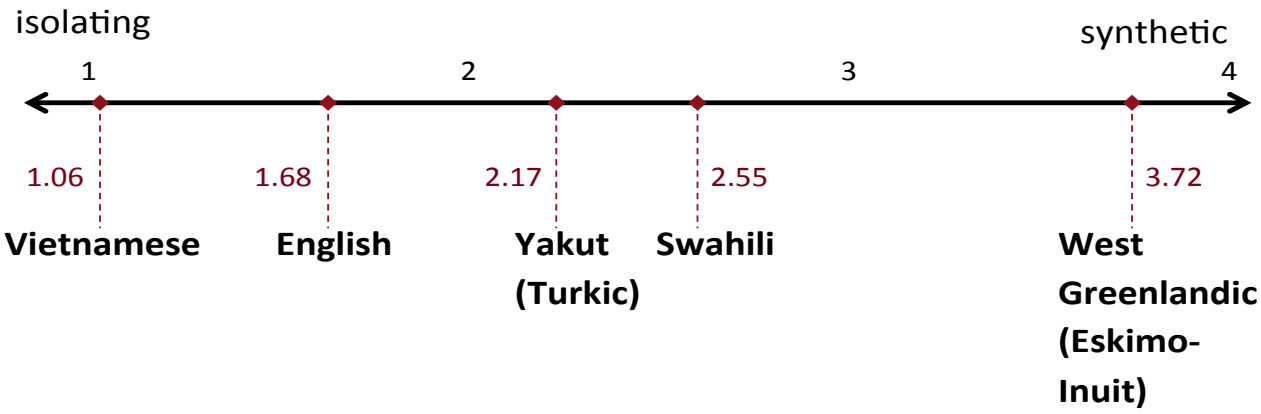
- MT is how NLP got started!

- Think about the domains of language discussed so far in this course. How do they differ from one language to another?
  - Morphology
  - Syntax
  - Semantics
  - Pragmatics and discourse

Morphological Differences

- Morpheme: "Minimal meaningful unit of a language"
  Word = Morpheme + Morpheme + Morpheme +…
  reading = read + ing;

- Number of morphemes per word



Source: Dan Jurafsky (MT lecture)

Morphological Differences

- Cantonese

"He said this was the biggest building in the whole country"

keui wa chyuhn gwok jeui daaih gaan nguk haih li gaan!
he say entire country most big bldg house is this bldg

Each word in this sentence has one morpheme (and one syllable)

- Turkish

uygarlaştıramadıklarımızdanmışsınızcasına

uygar+laş+tır+ama+dık+lar+ımız+dan+mış+sınız+casına

Behaving as if you are among those whom we could not cause to become civilized

Source: Dan Jurafsky (MT lecture)

Lexical Differences

- Word boundaries

  Boundaries between words not marked in some languages

  - Chinese, Japanese, Thai

- Lexical gaps

  One-to-one mapping between lexical items is rare

  - Chinese have no term for “brother”, “sister”, “aunt”, “uncle”…

- Mapping of words to phrases

  Computer science => informatique (French)

Syntactic Differences

Word order differences

- English/German/French/Mandarin: **subject-verb-object (SVO)**
- Hindi/Japanese: **subject-object-verb (SOV)**
- Irish, Classical Arabic: ***verb*-subject-object (VSO)**

English (SVO): *He adores listening to music*
Japanese (SOV): *He music to listening adores*

Semantic Differences

How languages distinguish spatial relations

- Satellite-framed languages:
  - direction of motion is marked on the satellite (particle, prepositional/adverbial phrases)
  - E.g., English, Chinese
- Verb-framed languages:
  - direction of motion is marked on the verb
  - E.g., Spanish, Japanese, French

English: The bottle **floated** out
Spanish: The bottle exited **floating**

Discourse & Pragmatic Differences

Usage of pronouns, discourse markers, grammatical politeness

- **Cold** language: References to actors are not explicit, e.g., Chinese, Japanese
- **Hot** language: References to actors are explicit, e.g., *English*

飓风丽塔已经减弱为第三级飓风，
Rita weakened and was downgraded to a Category 3 storm;

ø 迫近美国德课萨斯州和路易斯安那州，
[Rita/it/the storm] is moving close to Texas and Louisiana;

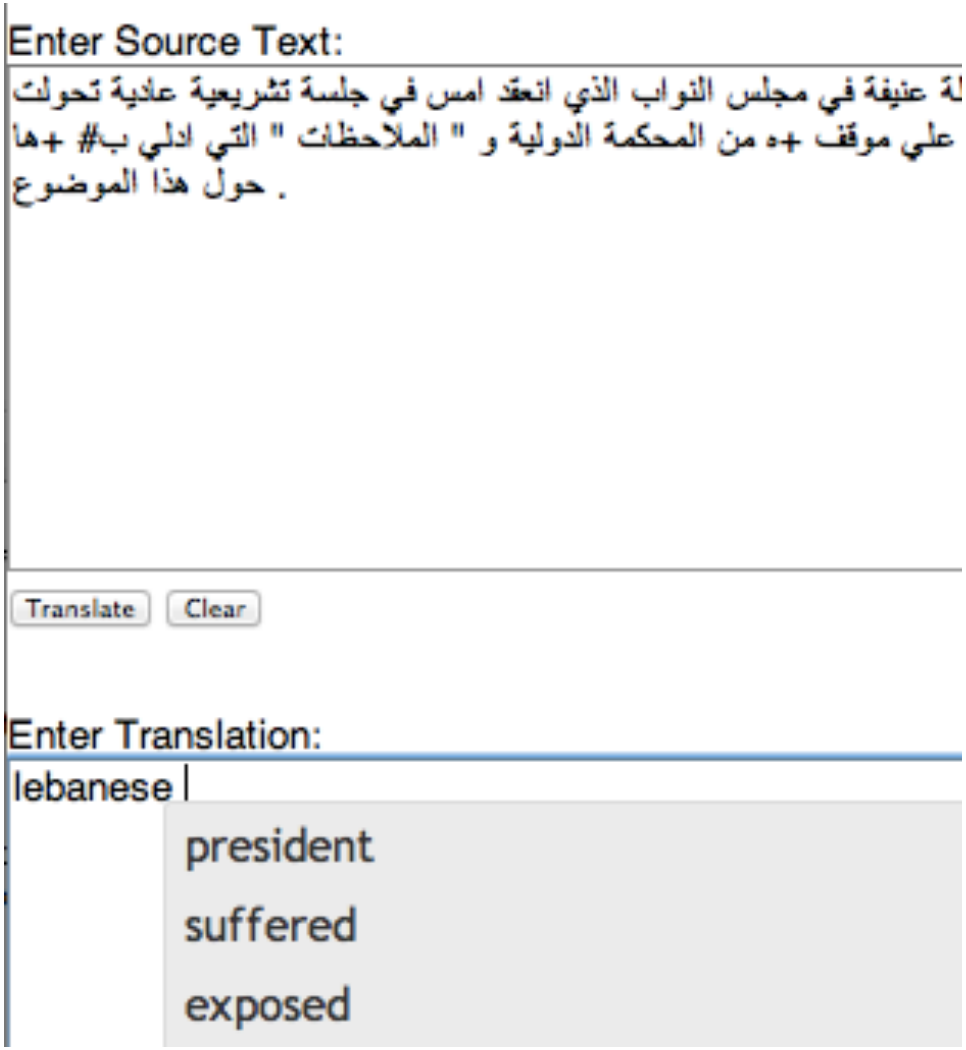
Usage Scenarios for MT

1. Fully automatic high-quality MT (FAHQMT)
   This dream still seems a distant goal!

2. Author-initiated high quality translation
   MT with human post-editing or MT as a translation aid is clearly growing ... but remains painful
   E.g., *MateCat or LiLT*
   https://lilt.com/

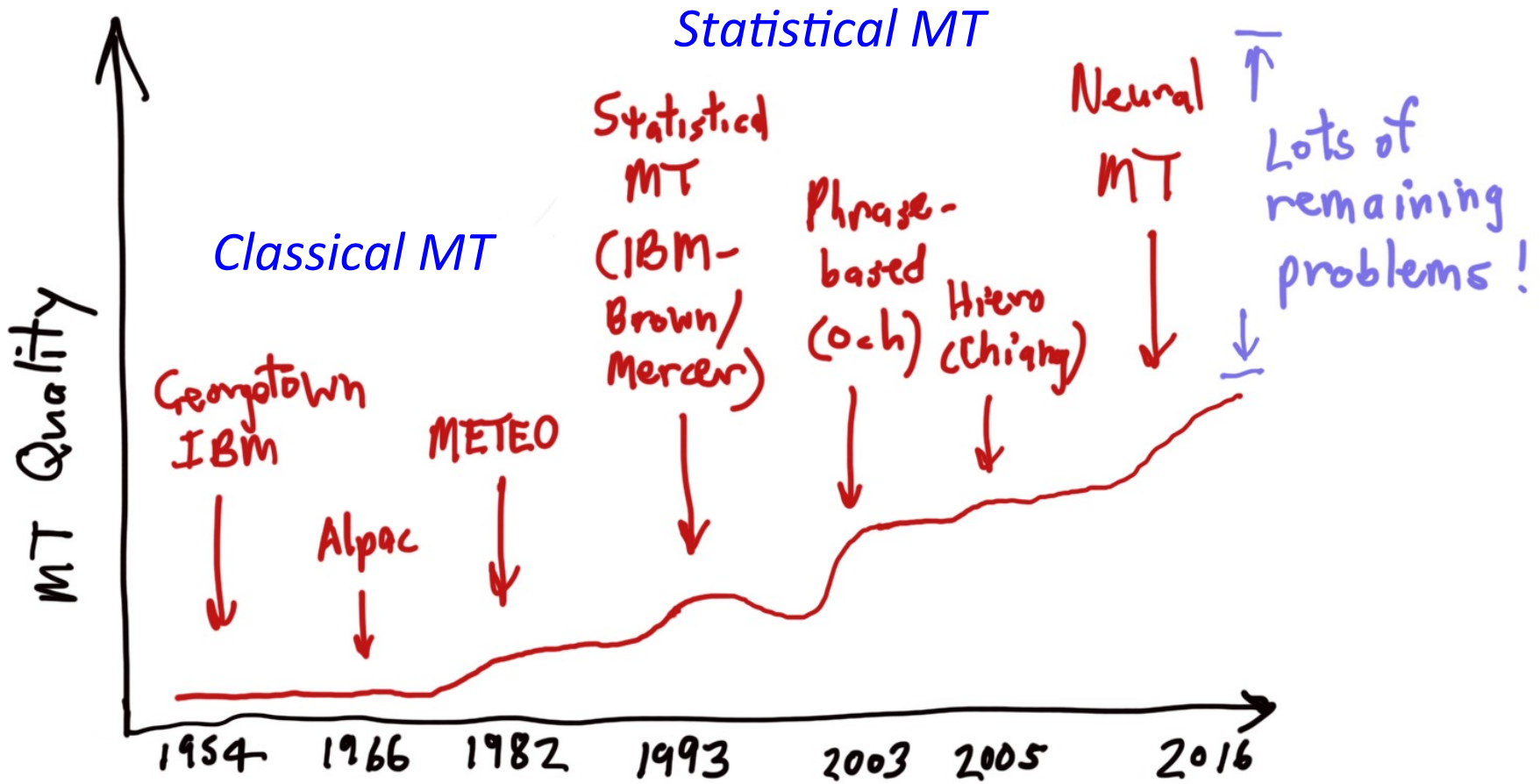
3. User-or platform-initiated low quality translation
   The current mainstay of MT
   Google Translate, Bing Translator

Helping Human Translators



Source: Dan Jurafsky (MT lecture)

Progress in MT

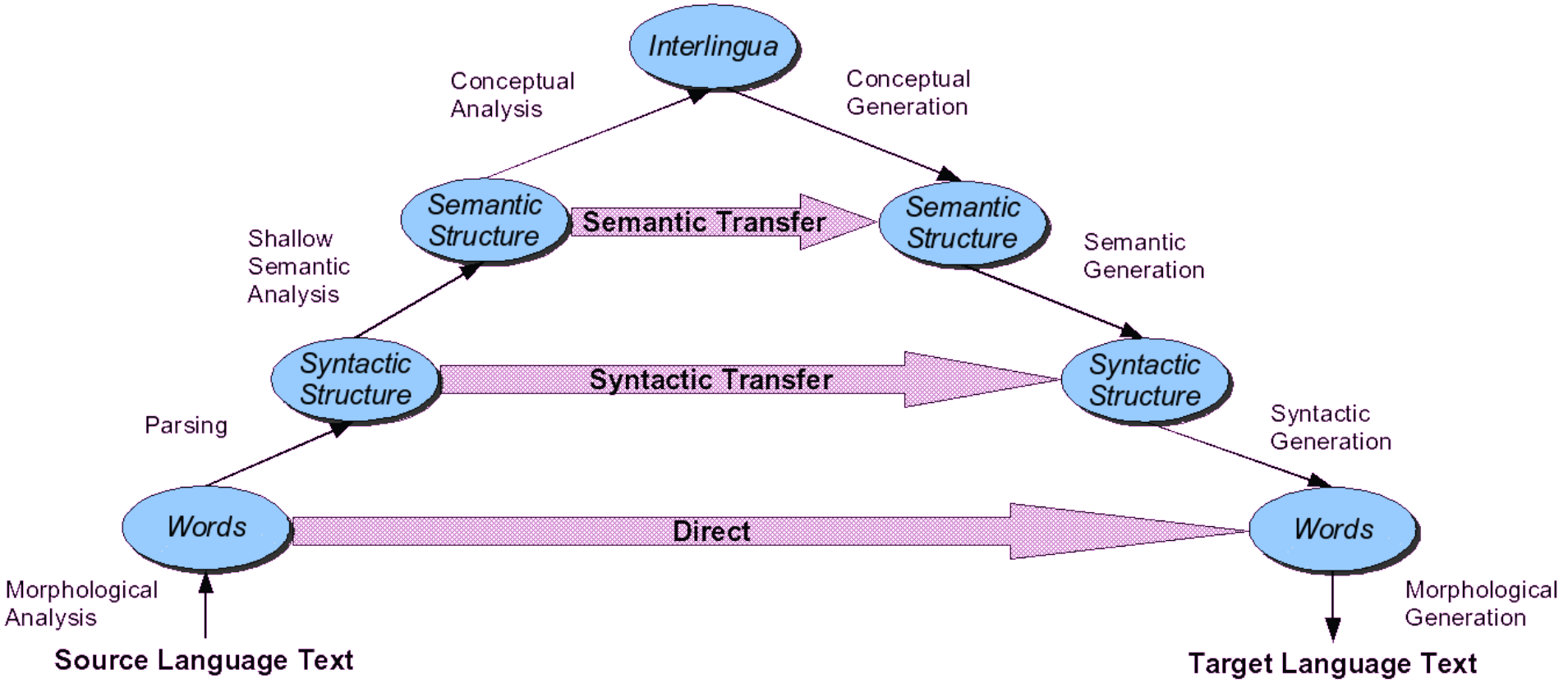


Source: Luong, Cho, Manning (ACL tutorial 2016)

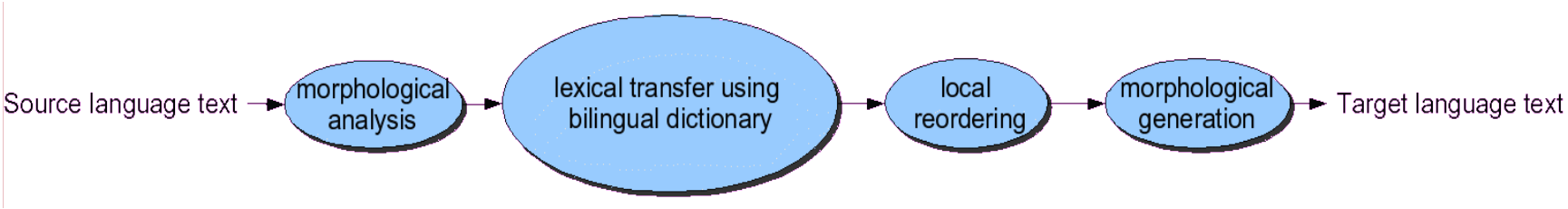
Classical Machine Translation

- Linguists write hand-crafted rules to account for language divergences.
- 3 classical methods for MT:
    - Direct
    - Transfer
    - Interlingua

The Vauquois Triangle



Direct Machine Translation



- Proceed word-by-word through text
- Very little analysis (only morphology) of the source/ target text (no syntactic or semantic analysis)
- Relies on a large bilingual dictionary. For each word in the source language, the dictionary *specifies a set of rules* for translating that word
- After the words are translated, simple reordering rules are applied (e.g., move adjectives after nouns when translating from English to French)

Direct Machine Translation

An example of a set of direct translation rules

Rules for translating much or many into Russian:

if preceding word is *how* **return** *skol'ko*
else if preceding word is *as* **return** *stol'ko zhe*
else if word is *much*
    if preceding word is *very* **return** nil
    else if following word is a noun **return** *mnogo*
else (word is many)
    if preceding word is a preposition and following word is noun **return** *mnogii*
    else return *mnogo*

Source: J & M (Ch 25); originally from Panov (1960)

Direct Machine Translation

A direct translation example

Input:	Mary didn't slap the green witch
After 1: Morphology	Mary DO-PAST not slap the green witch
After 2: Lexical Transfer	Maria PAST no dar una bofetada a la verde bruja
After 3: Local reordering	Maria no dar PAST una bofetada a la bruja verde
After 4: Morphology	Maria no dió una bofetada a la bruja verde

Source: J & M (Ch 25); originally from Panov (1960)

Problems with Direct Translation

- Lack of any structural analysis causes several problems

  Difficult or impossible to capture long-range reorderings

  English: Sources said that IBM bought Lotus yesterday **[SVO]**
  Japanese: Sources yesterday IBM Lotus bought that said **[SOV]**

  Words are translated without disambiguation of their syntactic role, e.g., that can be a complementizer or a determiner, and will often be translated differently for these two cases

  They said that ...
  They like that ice-cream

Transfer-based Translation

Three phases in translation:

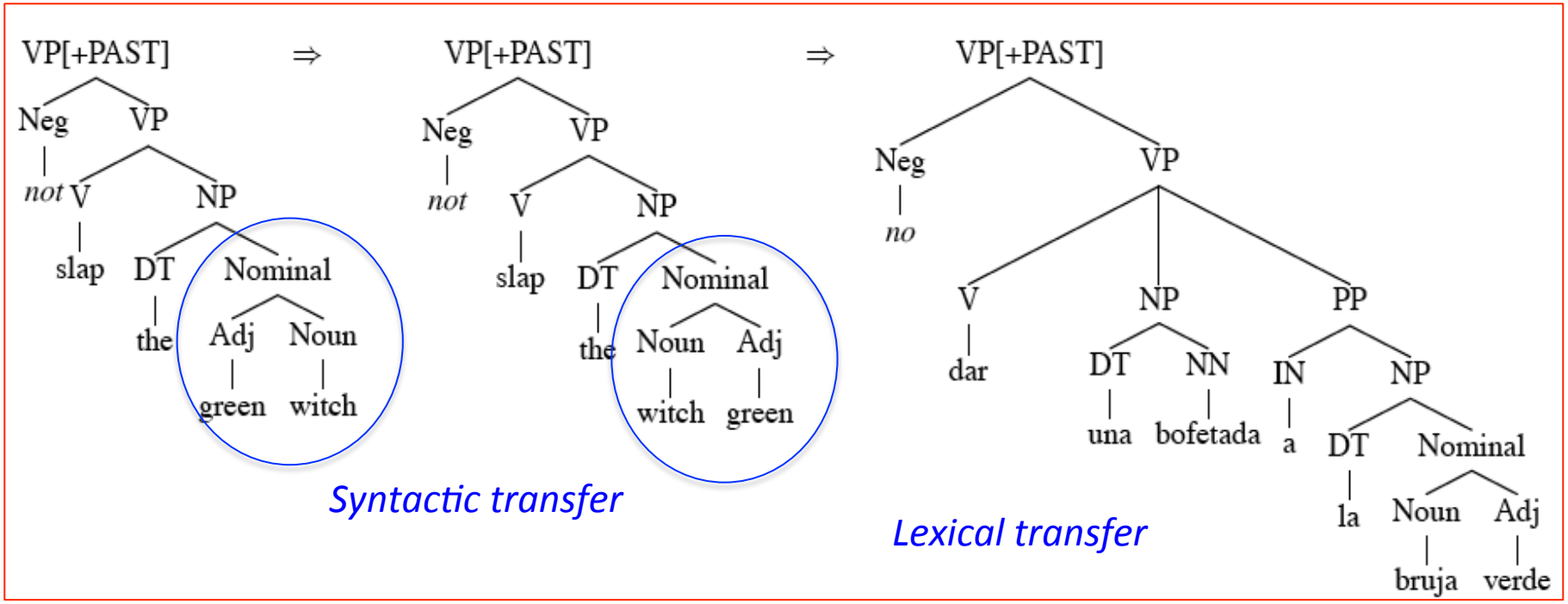
1. Analysis: Analyze the source language sentence; for example, build a syntactic analysis of the sentence.

2. Transfer: Convert the source-language parse tree to a target-language parse tree by considering the differences between the syntactic structures of the two languages.

3. Generation: Convert the target-language parse tree to an output sentence.

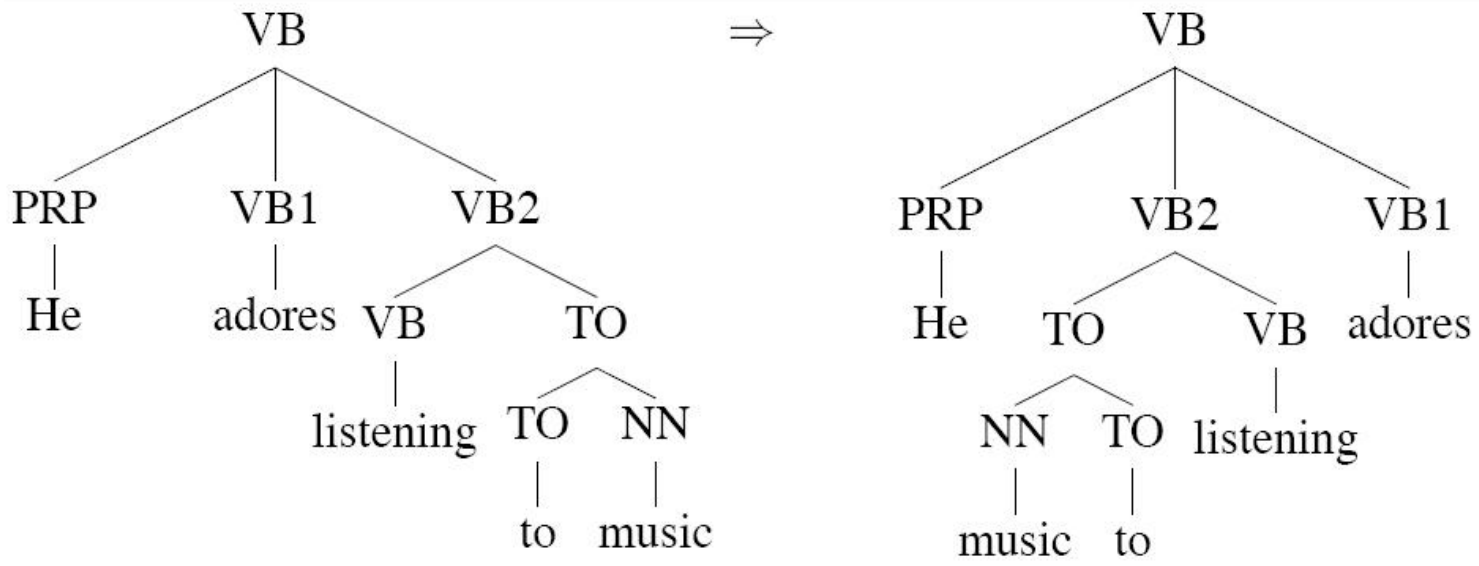
Transfer Rules

English ⇒ Spanish



Transfer Rules

English (SVO) Japanese (SOV)



Transfer-based systems also need lexical transfer rules like direct MT.

Systran system is a classic example of commercial transfer-based MT.

Interlingua-Based Translation

Two phases in translation:

1. Interpretation: Analyzes the source language sentence into a (language-independent) ***meaning representation***, i.e., interlingua.

2. Generation: Convert the meaning representation into a target language sentence.

Event-based Representation for Interlingua

Interlingua for *"Mary did not slap the green witch"*

$$
\begin{bmatrix}
\text{EVENT} & \text{SLAPPING} \\
\text{AGENT} & \text{MARY} \\
\text{TENSE} & \text{PAST} \\
\text{POLARITY} & \text{NEGATIVE} \\
\text{THEME} & \begin{bmatrix}
\text{WITCH} & \\
\text{DEFINITENESS} & \text{DEF} \\
\text{ATTRIBUTES} & \begin{bmatrix} \text{HAS-COLOR} & \text{GREEN} \end{bmatrix}
\end{bmatrix}
\end{bmatrix}
$$

Use other NLP tools to analyze the source sentence

Interlingua-Based Translation

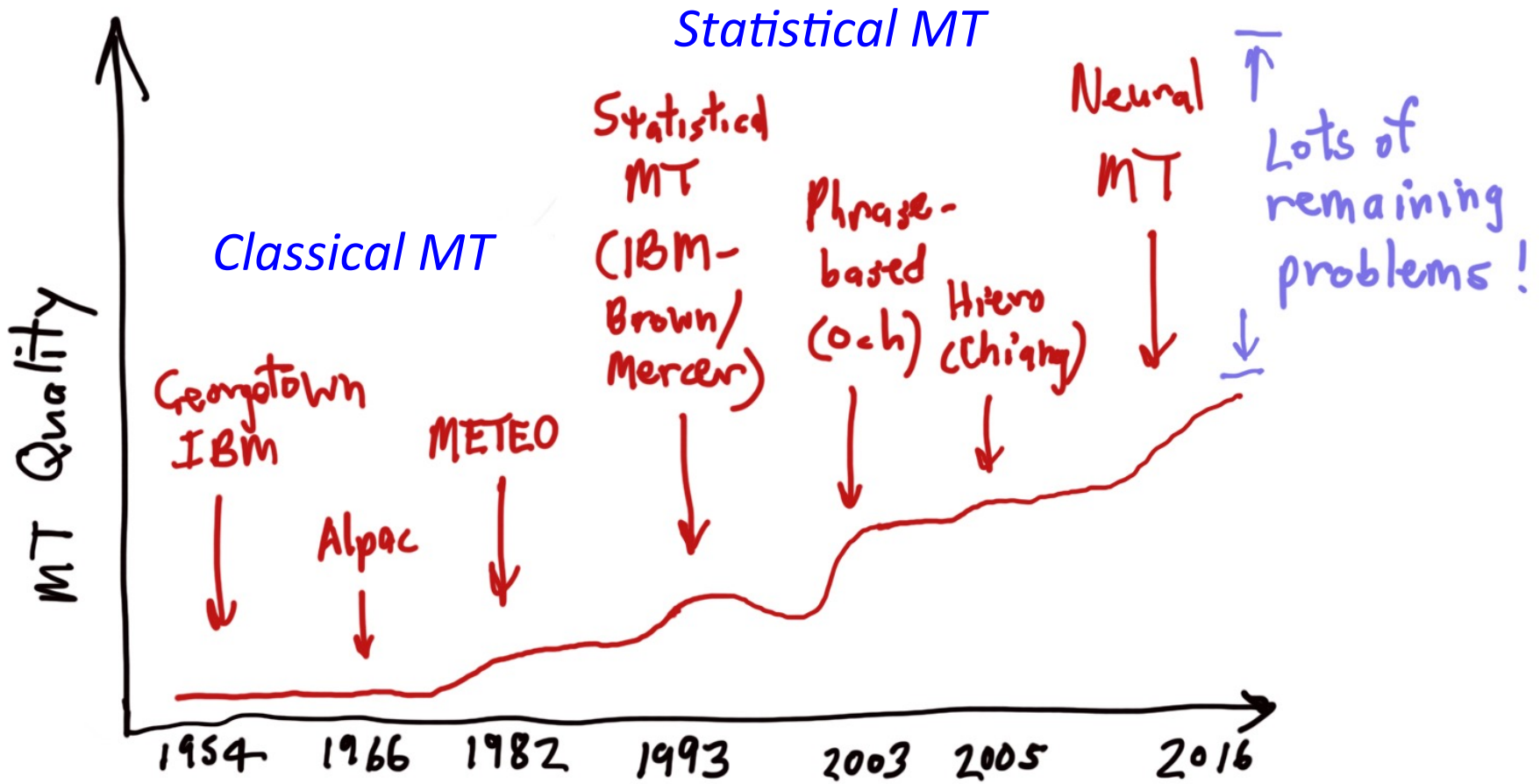
- Advantage: If we want to build a translation system that translates between *n* languages, we need to develop *n* analysis and *n* generation systems. With a transfer based system, we'd need to develop $O(n^2)$ sets of translation rules.

- Disadvantage: What would a language-independent representation look like? Different languages break down concepts in quite different ways:

  Japanese has two words for brother: one for an elder brother, one for a younger brother

  German has two words for wall: one for an internal wall, one for a wall that is outside

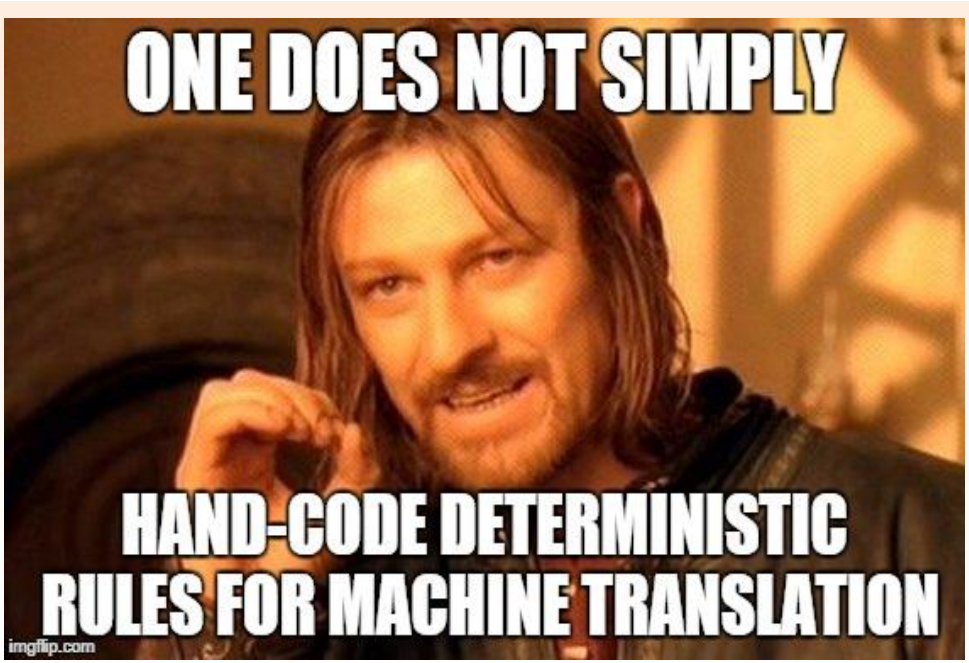
Progress in MT



Source: Luong, Cho, Manning (ACL tutorial 2016)

Problems with Rule-based Systems

- Writing rules requires a lot of human effort and knowledge



"Every time I fire a linguist, the performance of the speech recognizer goes up" Frederick Jelinek

Importance of Rules

- We should not undermine the importance of rules.
- Statistical MT relies on huge bilingual corpora, which may not be available for restricted domains (e.g., dialects, instructions).

Statistical Machine Translation (SMT)

Statistical Machine Translation

- Parallel corpora (aka bilingual, bitext) are available in several language pairs.

  Contains the "same" text in two or more languages

- Basic idea: use a parallel corpus as a training set of translation examples and learn a translation model.

- **E.g., Canadian Hansard** – parliamentary debates in English and French

E: *Canada should therefore drop any reference to any system other than the metric system in ads, on signs, and on packaging. The petitioners are also calling for containers to be standardized to the metric system in units of 100 grams or 100 millilitres.*

F: *Le Canada devrait donc abandonner toute référence à un système autre que le système métrique dans la publicité, l'affichage et sur les contenants. Les pétitionnaires demandent aussi l'uniformisation des contenants au système métrique par tranche de 100 grammes ou de 100 millilitres.*

Ms. Ève Péclet, 41st Parliament, #232

Statistical Machine Translation

Intuition comes from the **impossibility** of perfect translation

- Perfect translation is impossible in most cases
- Translate Hebrew *adonai roi* ("the lord is my shepherd") to language/culture without sheep or shepherds
- Two options:
  - Something **fluent** and understandable, but not faithful:
    "The Lord will look after me!"
  - Something **faithful**, but not fluent or natural
    "The Lord is for me like somebody who looks after animals with cotton-like hair!"

One needs to compromise between fluency and faithfulness

Statistical MT: Faithfulness and Fluency Formalized!

- Given a foreign (French) sentence ***F***, find an English sentence

$$\hat{E} = \underset{E \in English}{\operatorname{argmax}} P(E \mid F)$$

$$= \underset{E \in English}{\operatorname{argmax}} \frac{P(F \mid E)P(E)}{P(F)}$$

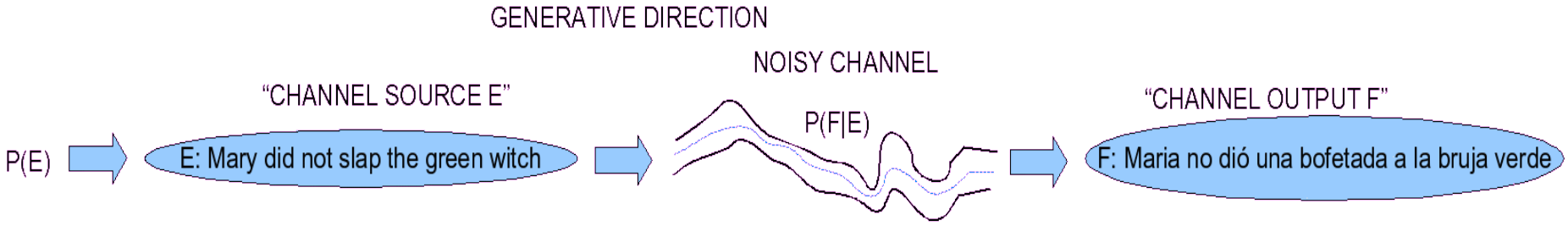
$$= \underset{E \in English}{\operatorname{argmax}} P(F \mid E)P(E)$$

Decoder (Beam search)

Translation model (faithful)

Language model (fluency)

The Noisy Channel Model for MT



$$\hat{E} = \underset{E \in English}{\operatorname{argmax}} P(E \mid F)$$

$$= \underset{E \in English}{\operatorname{argmax}} \frac{P(F \mid E)P(E)}{P(F)}$$

$$= \underset{E \in English}{\operatorname{argmax}} P(F \mid E)P(E)$$

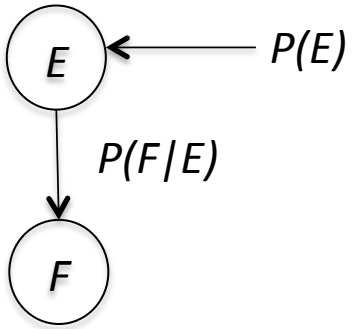
Decoder

Translation model (faithful)

Language model (fluency)

- Generative model

P(E|F)?



Fluency: P(E)

We need a metric that ranks this sentence

That car almost crash to me!

as less fluent than this one:

That car almost hit me.!

- Answer: language models

  P(me|almost, hit) > P(to|almost, crash)

- Advantage: this is monolingual knowledge.

Faithfulness: P(F|E)

- **Spanish:**

  Maria no dió una bofetada a la bruja verde

- **English candidate translations:**

  Mary didn't slap the green witch
  Mary not give a slap to the witch green
  The green witch didn't slap Mary
  Mary slapped the green witch

- More faithful translations will be composed of phrases that are *high probability* translations

  How often was "slapped" translated as "dió una bofetada" in a large **bitext** (parallel English-Spanish corpus)

Computational Tasks in Noisy Channel Model for MT

- Language Model: given E, compute P(E)
  good English string → high P(E)
  random word sequence → low P(E)

- Translation Model: given (F,E) compute P(F | E)
  (F,E) look like translations → high P(F | E)
  (F.E) don't look like translations → low P(F | E)

- Decoding Algorithm: given LM, TM, and F, find Ê
  Find translation E that maximizes P(E) * P(F | E)

Computing P(F|E)

- Goal is to compute P(F|E) from a bitext (E,F) corpus
- Three options:
  - Consider **sentence-pairs** (E,F) to compute P(F|E)?
    => Sparse
  - Consider **words-pairs** (e_i, f_j) of the sentences and then take conditional independence assumption to compute P(F|E)?
    => Word alignment; Word-based MT
  - Consider **phrase-pairs** ($\mathbf{e}_i$, $\mathbf{f}_j$) to compute P(F|E)?
    => Phrasal alignment; Phrase-based MT

Word-based Translation Model
IBM Model 1

$$P(F \mid E) = \sum_{A} P(F, A \mid E)$$

$$= \sum_{A} P(F \mid E, A)\ \ P(A \mid E)$$

$$P(F \mid E, A) = \prod_{j=1}^{J} t(f_j, e_{a_j})$$

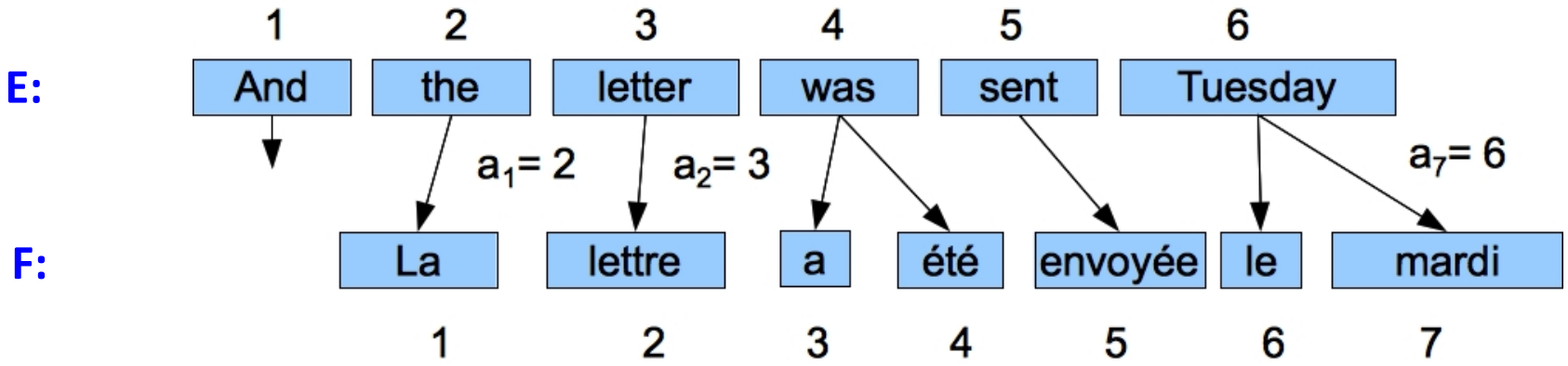
Translation probability

$$P(A \mid E) = \frac{\varepsilon}{(I+1)^J}$$

Uniform (known constant)

Word Alignment

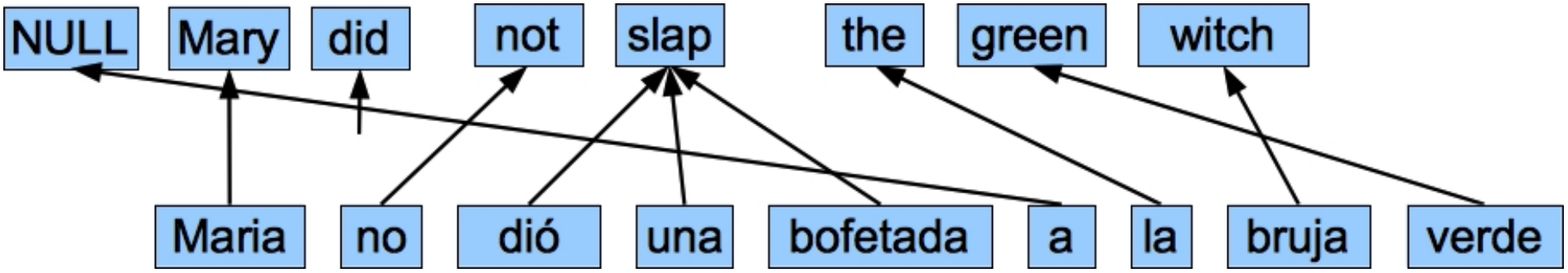
- A mapping between words in F and words in E



- Simplifying assumptions:
  - One-to-many (not many-to-one or many-to-many)
  - Each French word comes from exactly one English word
  - An alignment is thus a vector -- one cell for each French word
  - Alignment above A = [2, 3, 4, 4, 5, 6, 6]

One Addition: Spurious Words

- A foreign word that doesn't align with any word in the English sentence is called a spurious word.
- We model these by pretending they are generated by a NULL English word e_0 :



A = 1, 3, 4, 4, 4, 0, 5, 7, 6

Word Alignment and Translation Probability Computation

$$P(F \mid E) = \sum_{A} P(F, A \mid E)$$

$$= \sum_{A} P(F \mid E, A) \; P(A \mid E)$$

$$P(F \mid E, A) = \prod_{j=1}^{J} t(f_j, e_{a_j})$$

$$P(A \mid E) = \frac{\varepsilon}{(I+1)^J}$$

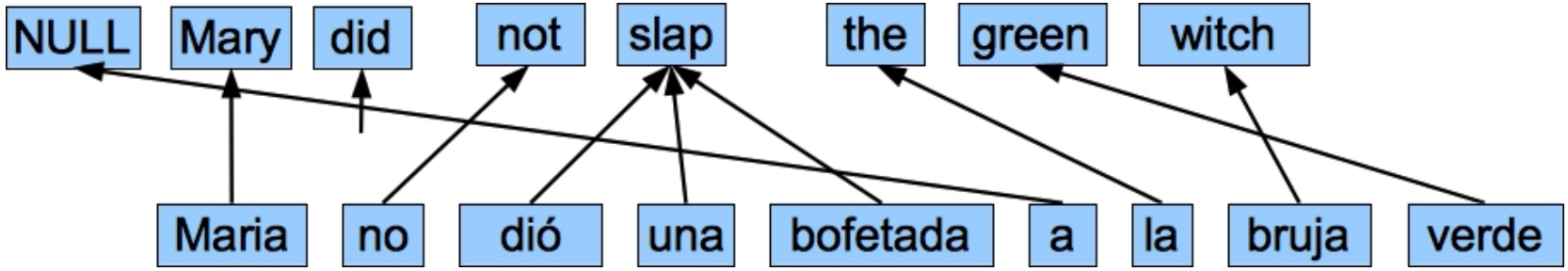
Given a pair of sentences ***(E, F)*** - one English, one French

- Learn the alignment ***A***, i.e., which English words align to which French words
- Learn the translation probability *t*

Word Alignment and Translation Probability Computation

But, *t* and *A* are dependent on each other!

- If we had known *A*, we could easily compute *t*



A = 1, 3, 4, 4, 4, 0, 5, 7, 6

t(verde, green) = n(green->verde) / n(green)

Word Alignment and Translation Probability Computation

But, *t* and *A* are dependent on each other!

- Similarly If we had known *t*, we could easily compute *A*

$$P(A|E,F) = \frac{P(A,F|E)}{\sum_A P(A,F|E)} = \frac{\prod_{j=1} t(f_j|e_{a_j})}{\sum_A \prod_{j=1}^{J} t(f_j|e_{a_j})}$$

Viterbi decoding is used to compute the best *A*

Word Alignment and Translation Probability Computation

- Unfortunately, we know neither *A* nor *t*

Solution: The **EM (Expectation-Maximization)** algorithm

- Basic idea: two steps to iterate over
  E Step: Assume *t* is known, compute expected *A*
  M Step: Re-compute *t* using the expected *A*

EM for Word Alignment and Translation Probability

1. Initialize *t* with uniform distribution
2. Repeat
   - **E Step:** Compute

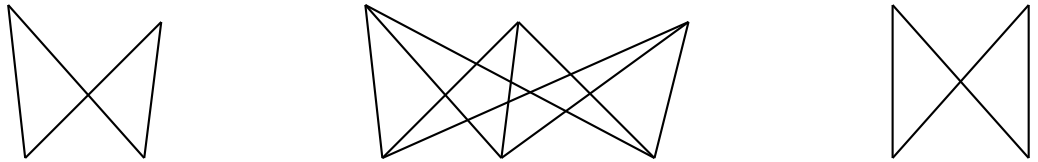
$$P(A|E,F) = \frac{P(A,F|E)}{\sum_A P(A,F|E)} = \frac{\prod_{j=1} t(f_j|e_{a_j})}{\sum_A \prod_{j=1}^{J} t(f_j|e_{a_j})}$$

   - **M Step:** Re-estimate *t* using *P(A|E, F)*

Until converge (i.e., parameters no longer change)

EM for Word Alignment and Translation Probability

… la maison … la maison bleue … la fleur …



… the house … the blue house … the flower …

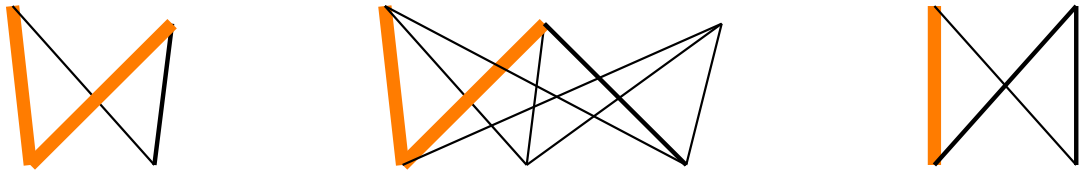
Initial stage:

All word alignments equally likely

All P(french-word | english-word) equally likely

EM for Word Alignment and Translation Probability

… la maison … la maison bleue … la fleur …

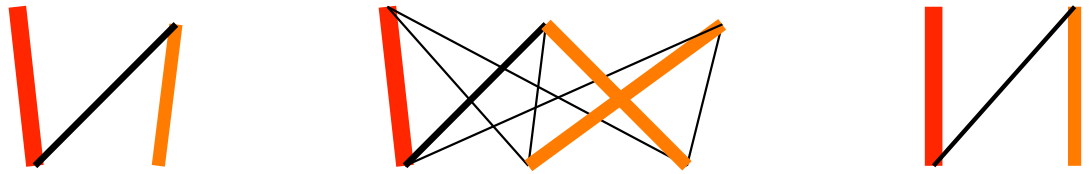


… the house … the blue house … the flower …

"la" and "the" observed to co-occur frequently, so P(la | the) is increased.

EM for Word Alignment and Translation Probability

… la maison … la maison bleue … la fleur …



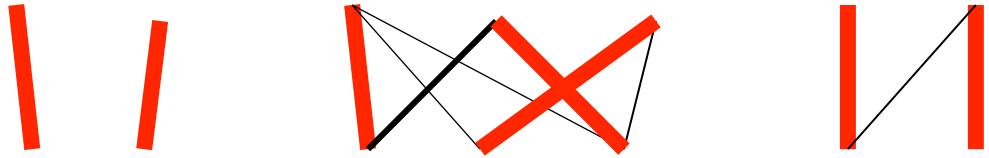
… the house … the blue house … the flower …

"house" co-occurs with both "la" and "maison",

- but P(maison | house) can be raised without limit, to 1.0
- while P(la | house) is limited because of "the" (Bayes property)

EM for Word Alignment and Translation Probability

… la maison … la maison bleue … la fleur …



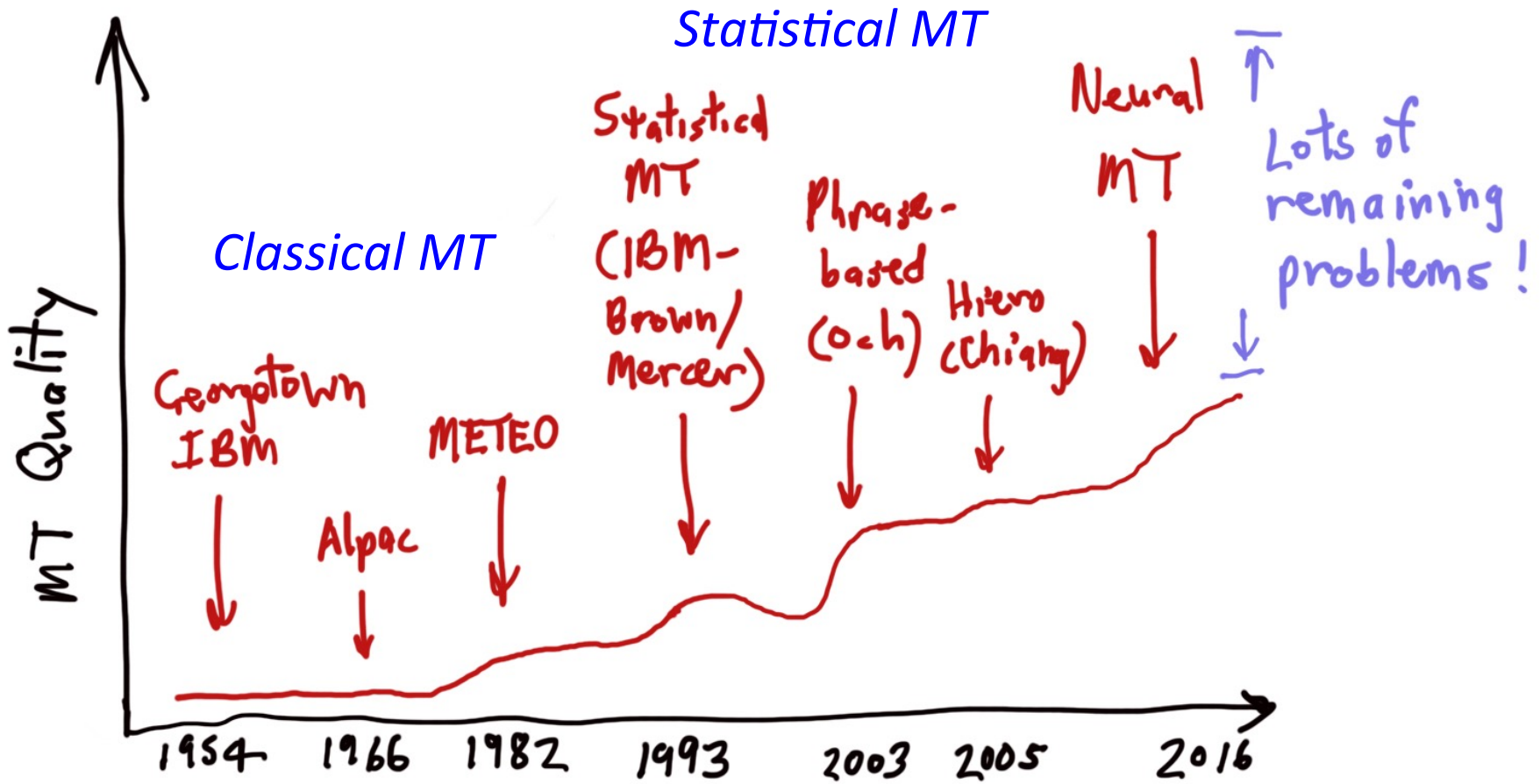
… the house … the blue house … the flower …

- settling down after another iteration

… la maison … la maison bleue … la fleur …

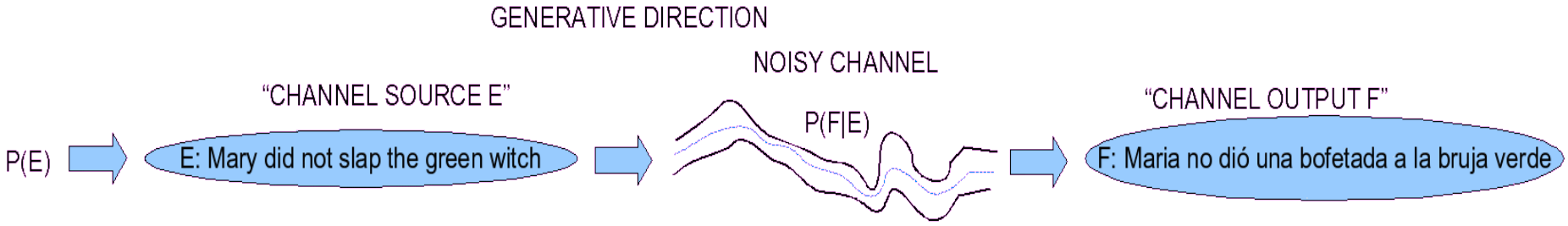
… the house … the blue house … the flower …

Progress in MT



Source: Luong, Cho, Manning (ACL tutorial 2016)

The Noisy Channel Model for MT



$$\hat{E} = \underset{E \in English}{\operatorname{argmax}} P(E \mid F)$$

$$= \underset{E \in English}{\operatorname{argmax}} \frac{P(F \mid E)P(E)}{P(F)}$$

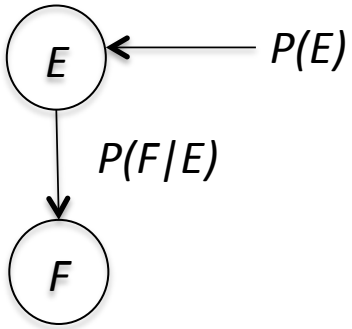
$$= \underset{E \in English}{\operatorname{argmax}} P(F \mid E)P(E)$$

Decoder

Translation model (faithful)

Language model (fluency)

- Generative model



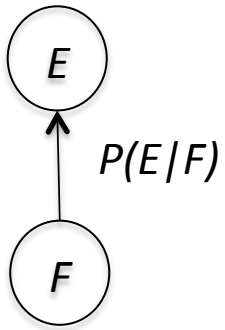
Log-linear Model for MT

- Instead of using the Bayes rule to compute P(E|F), we can model P(E|F) directly

- Discriminative model

weights *feature functions*

$$P(E|F) = \frac{\exp[\mathbf{w}^T \phi(E, F)]}{\sum_{E'} \exp[\mathbf{w}^T \phi(E', F)]}$$



- Feature functions:
  - language model
  - distortion model
  - P(F|E) translation model
  - P(E|F) reverse translation model
  - word penalty
  - phrase penalty

Problems with Traditional SMT

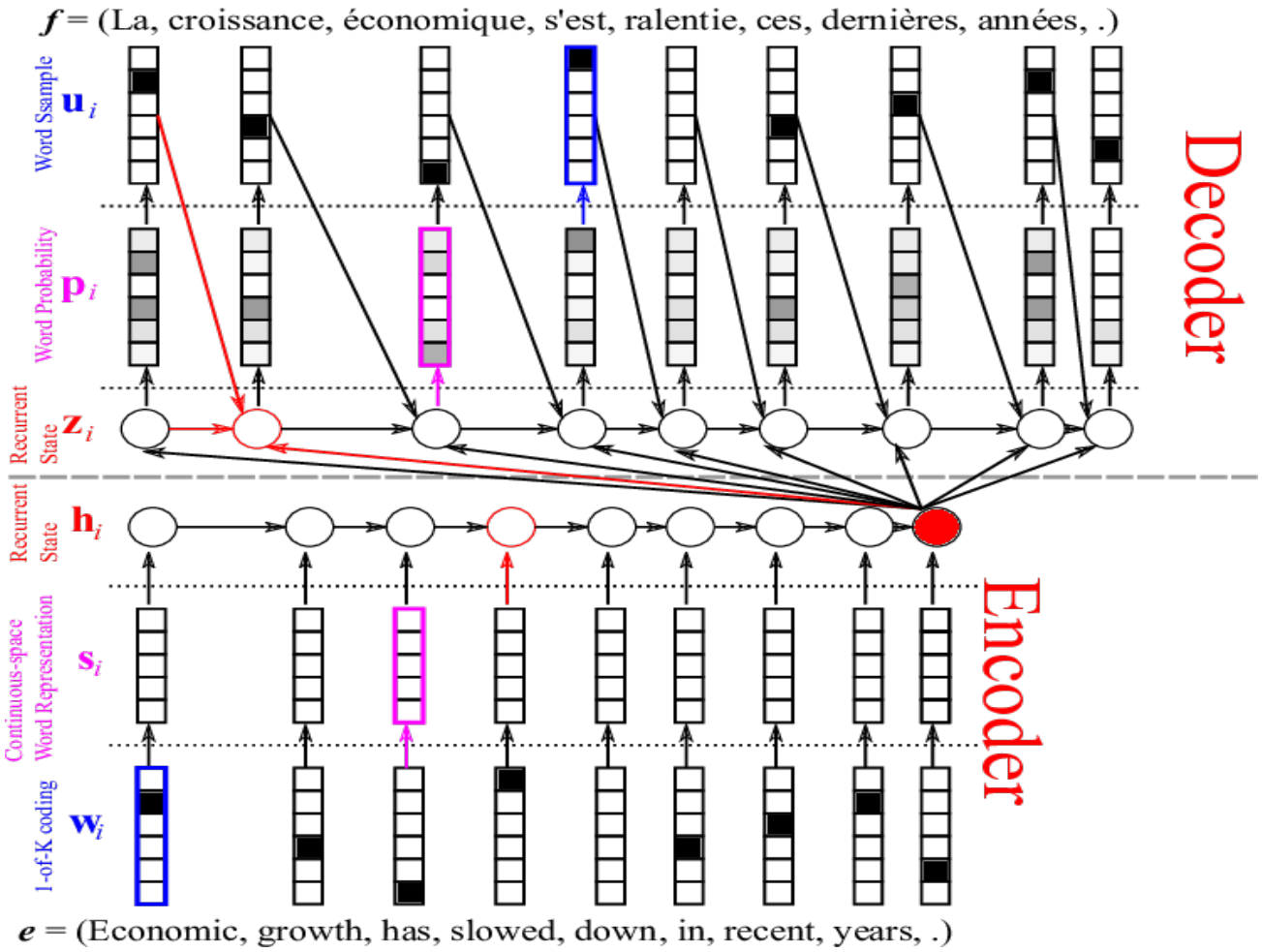
- Both noisy channel and log-linear models learn components (LM, TM, DM) separately and then put them together in decoder
- Each component is optimized with a different objective
- Discrete representation of words and phrases
- Context is limited

Neural MT

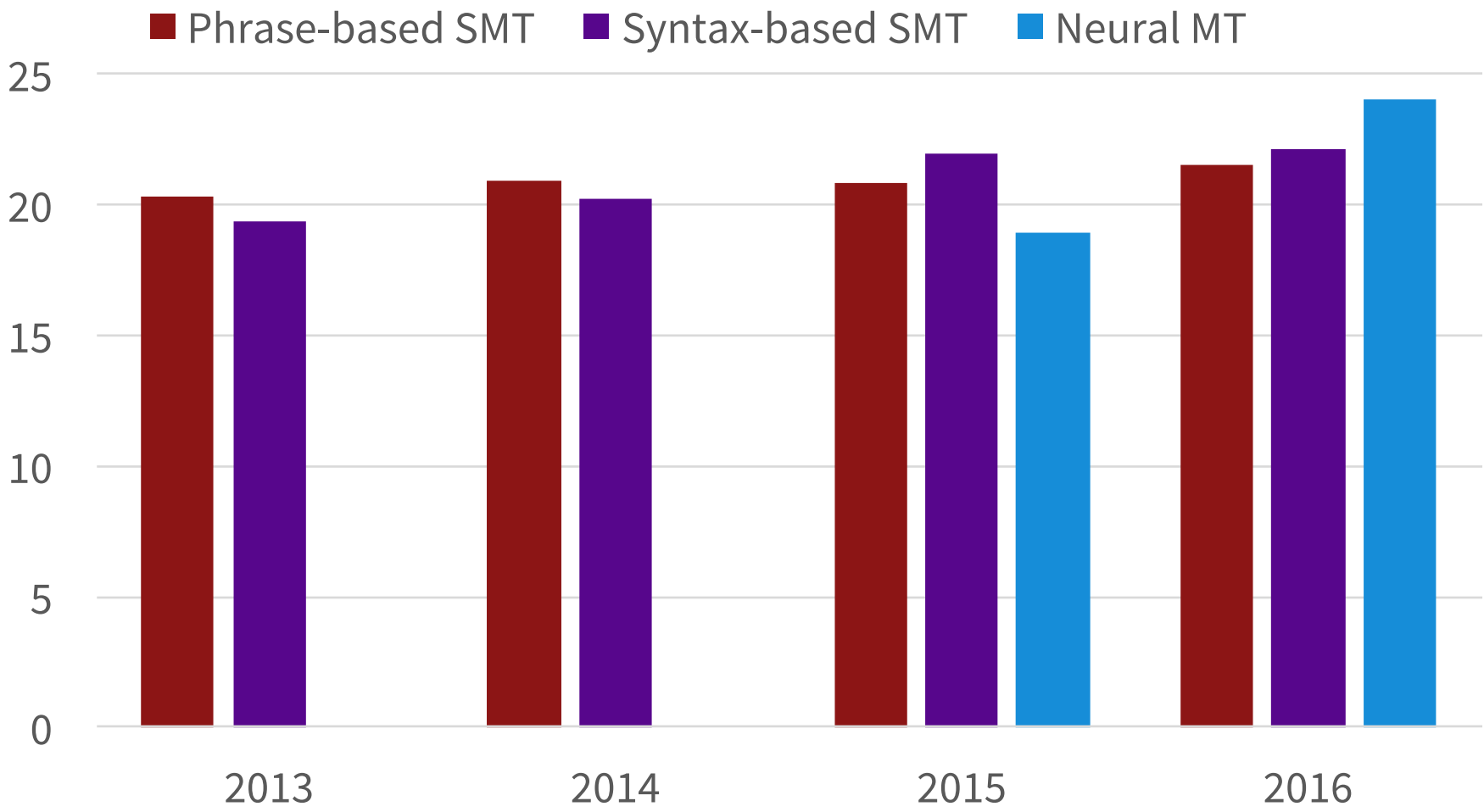
- New paradigm to MT; just came in 2015
- End-to-end training as a giant neural network model
  - Encoder
  - Decoder
- Distributed representation of words and phrases
- Allows longer context

Neural MT: http://104.131.78.120/

Neural MT


f = (La, croissance, économique, s'est, ralentie, ces, dernières, années, .)
Word Ssample
u_i
Word Probability
p_i
Recurrent State
z_i
Decoder
Recurrent State
h_i
Continuous-space Word Representation
s_i
1-of-K coding
w_i
Encoder
e = (Economic, growth, has, slowed, down, in, recent, years, .)

Recent Results



Thanks!

Questions?

Back up Slides

Machine Translation (MT)

- French <=> English

Obama et Romney prévoient de mener campagne dans les «swing states» à un rythme effréné pour les quatre derniers jours avant l'élection. L'Ohio se présente comme l'Etat le plus disputé du pays.

Obama and Romney plan to campaign in the "swing states" at a breakneck pace for the last four days before the election. The Ohio State presents itself as the most played country.

Example from Koehn (2016)

A Brief History of MT

- Machine Translation (MT) is how NLP got started!

- Early researchers in the 1950s were wildly optimistic.
- Georgetown-IBM experiment:

  A demonstration of Russian to English MT, featuring 6 translation rules and knowledge of around 250 words in the two languages.
- This resulted in substantial interest and funding for MT
- See this paper for details:

http://www.hutchinsweb.me.uk/AMTA-2004.pdf

A Brief History of MT

- Researchers in 1950s thought that with a little bit more work in engineering the rules and a more complete dictionary of words, they could develop a passable system.
- BUT They were wrong!!
- MT turned out to be a quite hard problem.
- The Automatic Language Processing Advisory Committee **(ALPAC)** report in 1966 criticized MT and its prospects

PopSci YouTube videos on the topic:

https://www.youtube.com/watch?v=5sLbWItc33I
https://www.youtube.com/watch?v=2ac41CO7Nr0

Statistical Machine Translation

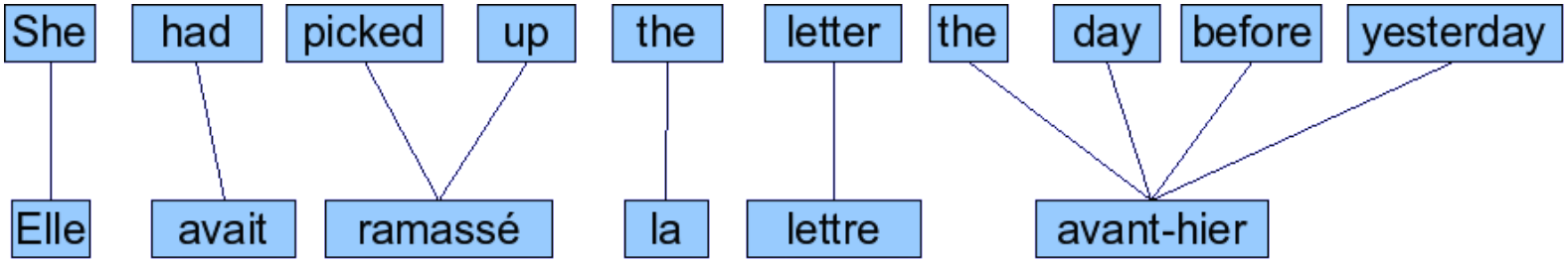
- Idea goes back to Warren Weaver (1949): suggested applying statistical and cryptanalytic techniques to translation.

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 (1949)

Alignments that don't Obey One-to-many Restriction

- Many to one:



- Many to many: