statistical machine translation

Part 3: Decoding & Evaluation

CSC401/2511 – Natural Language Computing – Spring 2019
Lecture 6 Frank Rudzicz and Chloé Pou-Prom
University of Toronto

Statistical Machine Translation

- Challenges to statistical machine translation
- Sentence alignment
- IBM model
- Phrase-based translation
- Decoding
- Evaluation

How to use the noisy channel

- How does this work?

$$E^* = \underset{E}{\operatorname{argmax}} P(F|E)P(E)$$

- $P(E)$ is a **language model** (e.g., *N*-gram) and encodes knowledge of word order.
- $P(F|E)$ is a **word-level translation model** that encodes only knowledge on an *unordered* word-by-word basis.

- **Combining** these models can give us **naturalness** and **fidelity**, respectively.

How to use the noisy channel

- Example from Koehn and Knight using only conditional likelihoods of Spanish words given English words.

- *Que hambre tengo yo*

  →

  What hunger have I $P(S|E) = 1.4E^{-5}$

  Hungry I am so $P(S|E) = 1.0E^{-6}$

  I am so hungry $P(S|E) = 1.0E^{-6}$

  Have I that hunger $P(S|E) = 2.0E^{-5}$

  …

UNIVERSITY OF TORONTO

How to use the noisy channel

- … and with the English language model

- *Que hambre tengo yo*
  →
  What hunger have I $P(S|E)P(E) = 1.4E^{-5} \times 1.0E^{-6}$
  Hungry I am so $P(S|E)P(E) = 1.0E^{-6} \times 1.4E^{-6}$
  I am so hungry $P(S|E)P(E) = 1.0E^{-6} \times 1.0E^{-4}$
  Have I that hunger $P(S|E)P(E) = 2.0E^{-5} \times 9.8E^{-7}$
  …

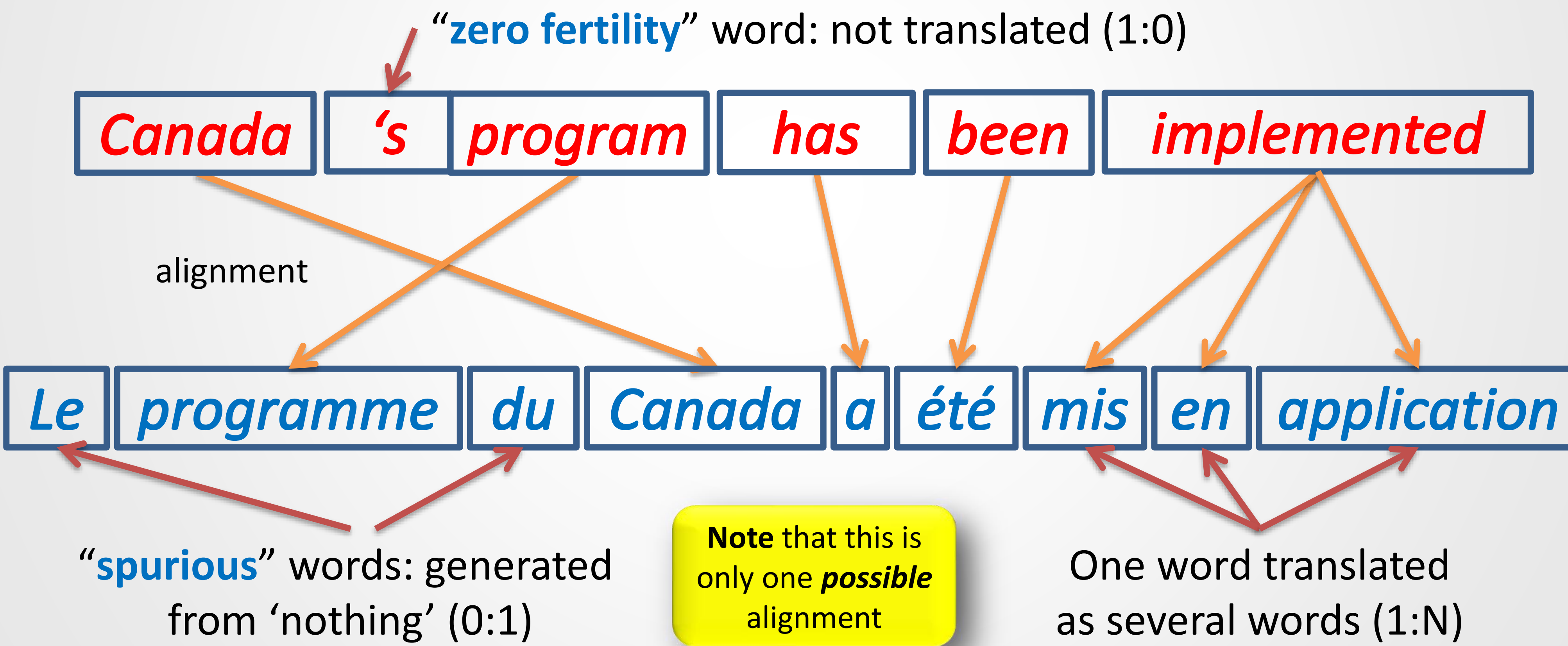
Sentence alignment

- We often need to align **sentences** before we can align **words**.

- We'll look at two broad classes of methods:
  1. Methods that only look at **sentence length**,
  2. Methods based on **lexical matches**, or "cognates".

Word alignment

- **Word alignments** can be 1:1, N:1, 1:N, 0:1,1:0,… E.g.,

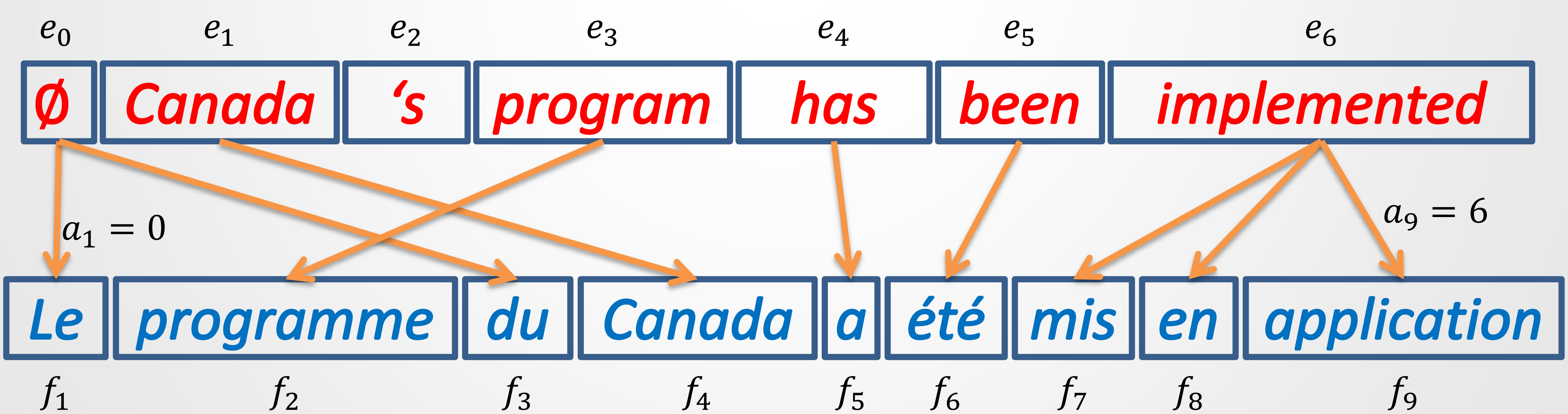


IBM models

IBM Model 1	lexical translation
IBM Model 2	adds absolute **re-ordering model**
IBM Model 3	adds **fertility model**

IBM Model 1: alignments

- An **alignment**, a, identifies **the English word** that 'produced' **the given French word at each index.**
  - $\boldsymbol{a} = \{\boldsymbol{a_1}, \dots, \boldsymbol{a_{L_F}}\}$ where $a_j \in \{0, \dots, L_E\}$
  - E.g., $\boldsymbol{a} = \{\boldsymbol{0, 3, 0, 1, 4, 5, 6, 6, 6}\}$



UNIVERSITY OF TORONTO

IBM-1: expectation-maximization

1. **Initialize** translation parameters $P(f|e)$ (e.g., randomly).

2. **Expectation**: Given the current $\theta_k = P(f|e)$, compute the **expected value** of $\boldsymbol{Count}(\boldsymbol{f}, \boldsymbol{e})$ for all words in training data $\mathcal{O}$.

3. **Maximization**: Given the expected value of $\boldsymbol{Count}(\boldsymbol{f}, \boldsymbol{e})$, compute the **maximum likelihood** estimate of $\theta_k = P(f|e)$

UNIVERSITY OF TORONTO

IBM-1: expectation-maximization

- First, we **initialize** our parameters, $\boldsymbol{\theta} = \boldsymbol{P}(\boldsymbol{f}|\boldsymbol{e})$.

- In the **Expectation** step, we compute **expected** counts:
  - $TCount(f, e)$: the total number of times e and f are aligned.
  - $Total(e)$: the total number of e.

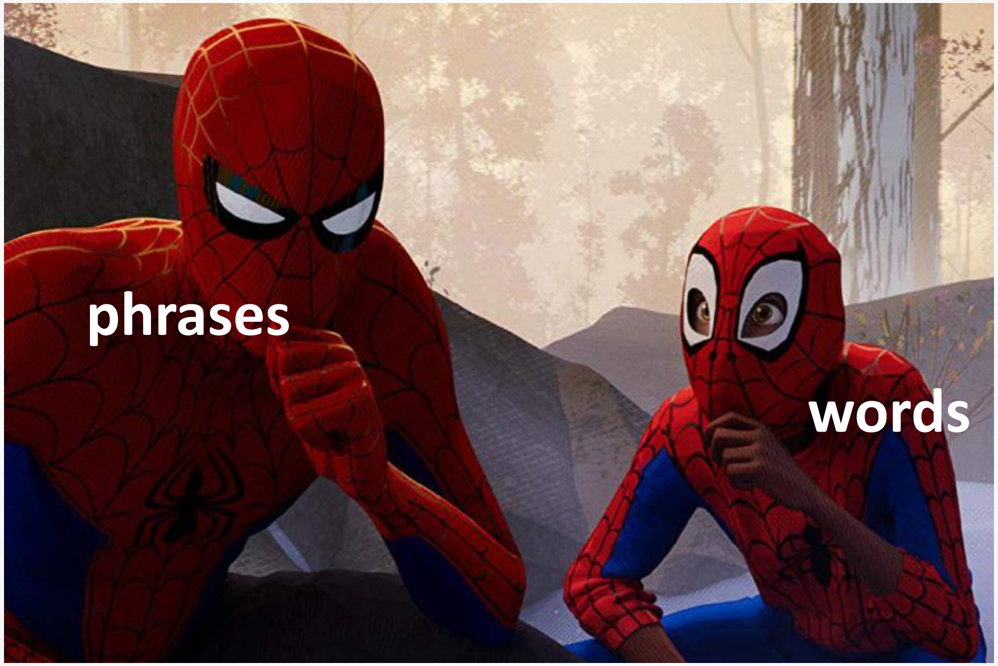
    This has to be done in steps by first computing $\boldsymbol{P}(\boldsymbol{F}, \boldsymbol{a}|\boldsymbol{E})$ *then* $\boldsymbol{P}(\boldsymbol{a}|\boldsymbol{F}, \boldsymbol{E})$

- In the **Maximization** step, we perform MLE with the expected counts.

IBM-1 EM

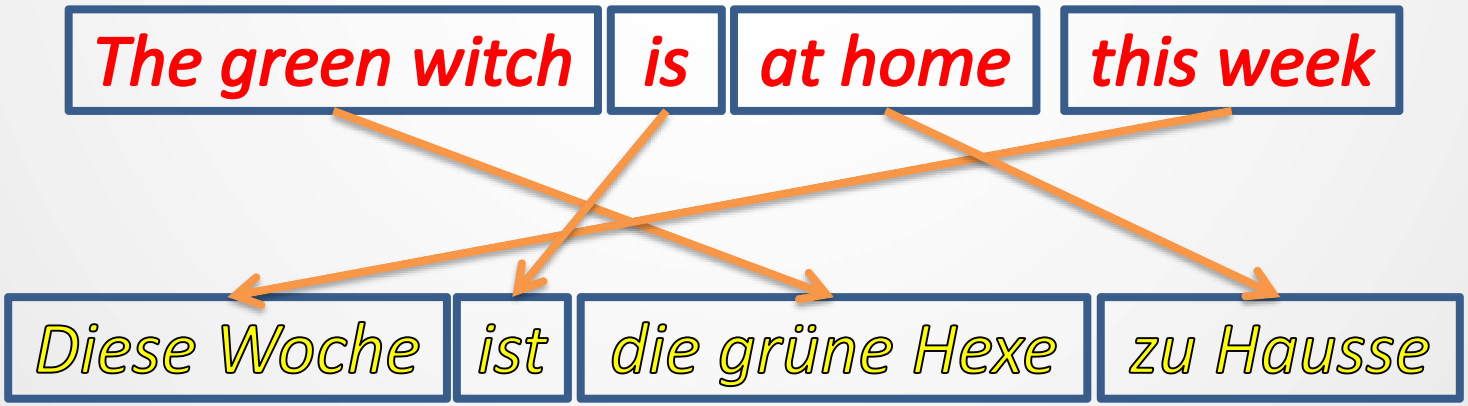
1. Initialize $P(f|e)$
2. Make grid of all possible alignments
3. Compute $P(F|a, E)$ → Products of $P(f|e)$
4. Compute $P(a|E, F)$ → Divide by sum of rows from step 3
5. Compute $TCount$→ Sum relevant probabilities from step 4
6. Compute $Total$ → Sum over rows from step 5
7. Compute $P(f|e) = \frac{TCount(f,e)}{Total(e)}$

UNIVERSITY OF TORONTO


phrases
words

Phrase-based statistical MT

- **Phrase-based** statistical MT involves segmenting sentences into contiguous blocks or segments.
  - Each phrase is probabilistically **translated**.
    e.g., $P(zu\ Hausse|at\ home)$
  - Each phrase is probabilistically **re-ordered**.



Phrase-based statistical MT

- Phrase-based SMT allows **many-to-many** word mappings.
- Larger context allows for some **disambiguation** that is not possible in word-based alignment.
  - E.g.,

$$P(coup|stroke)$$

vs.

$$P(coup\ de\ poing|punch) > P(coup\ de\ poing|stroke\ of\ fist)$$

$$P(coup\ d'oeil|glance) > P(coup\ d'oeil|stroke\ of\ eye)$$

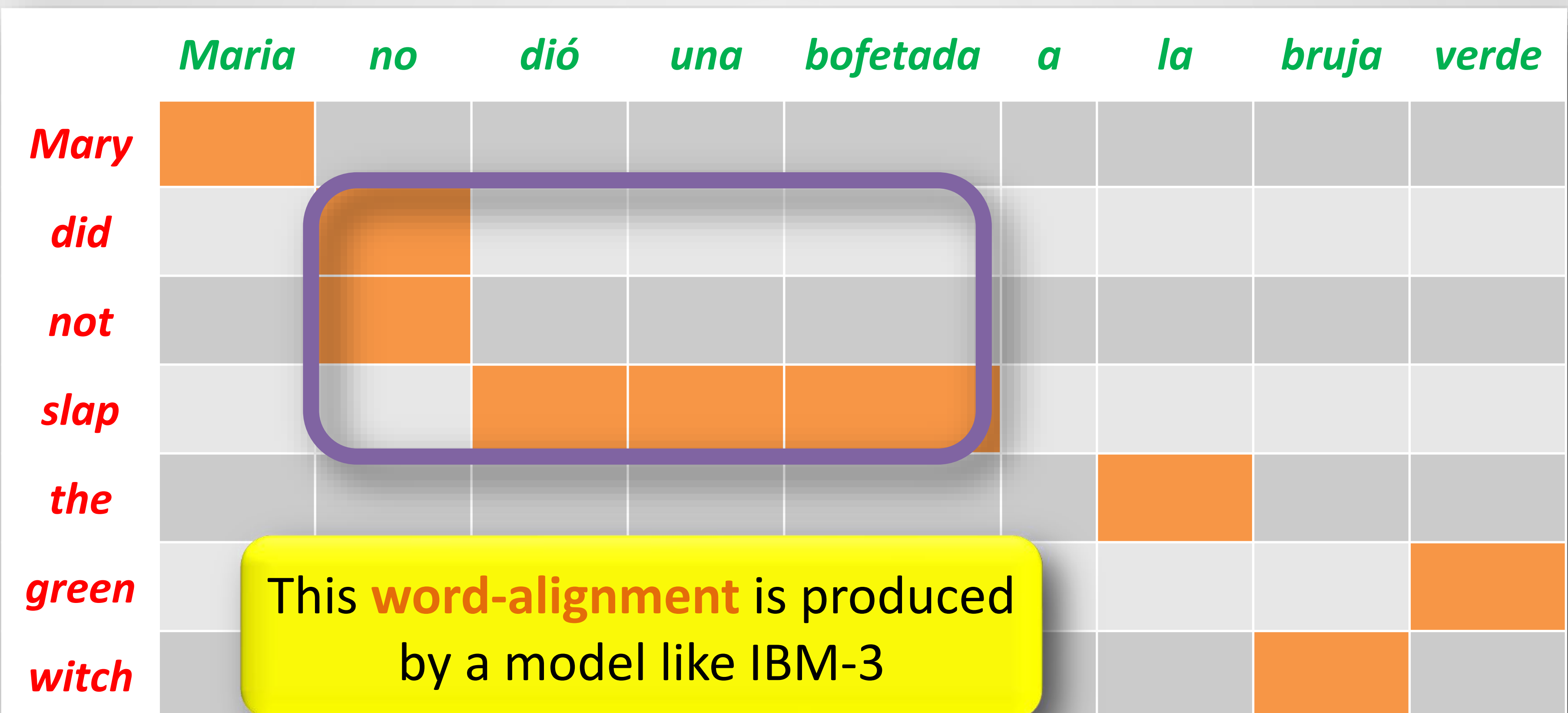
No context ☹

A tiny amount of context ☺

UNIVERSITY OF TORONTO

Learning phrase-translations

- Typically, we use **alignment templates** (Och *et al.*, 1999).
  - Start with a **word-alignment**, then build **phrases**.

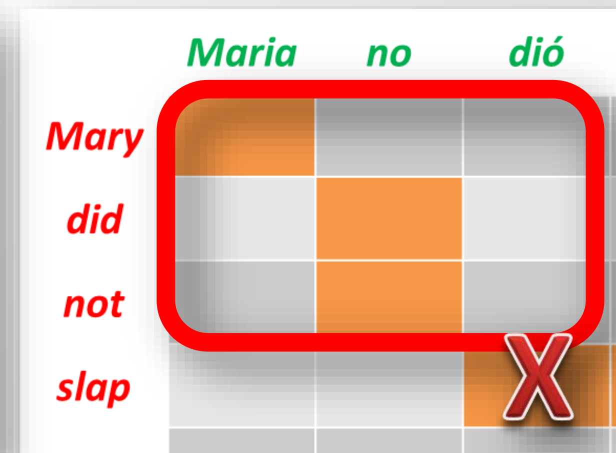


Learning phrase-translations

- A phrase alignment **must** contain **all** word alignments for each of its rows and columns.
  - Collect **all** phrase alignments that are consistent with the word alignment, e.g.

Consistent

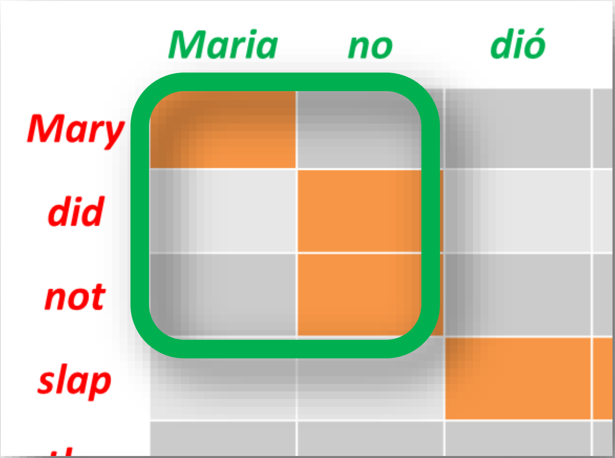
Inconsistent



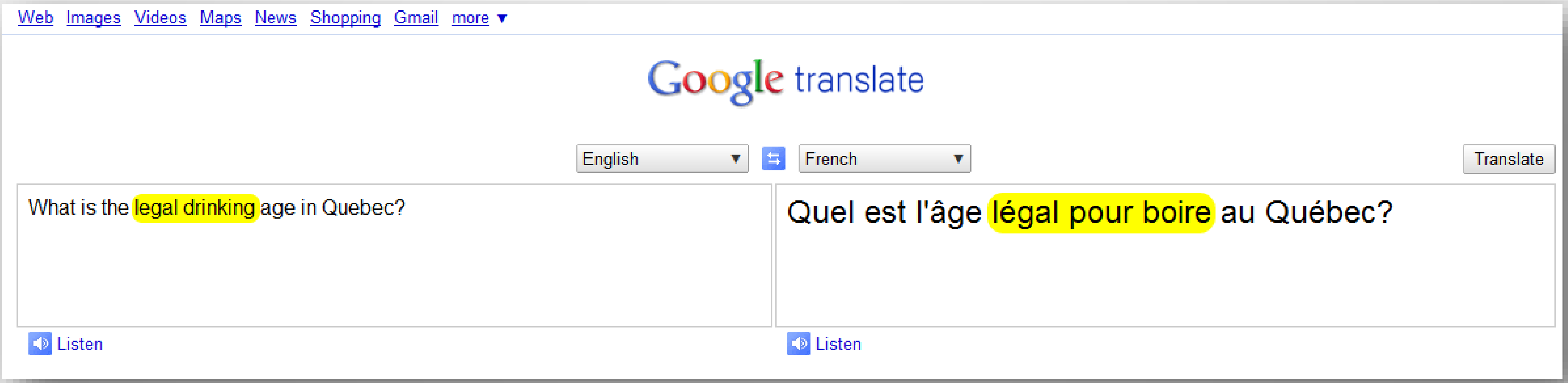
Inconsistent

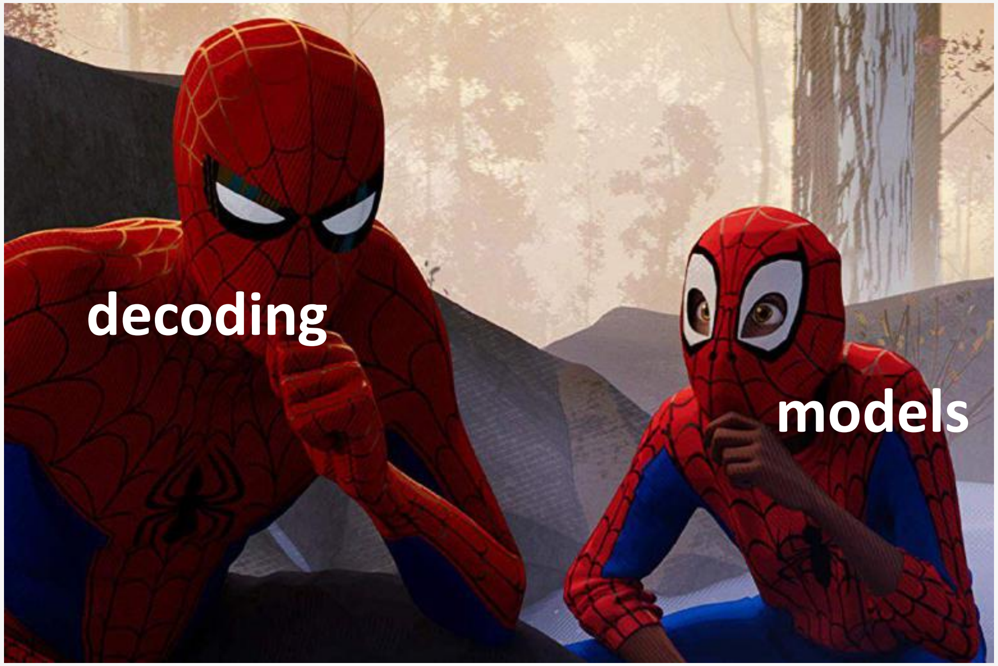
Learning phrase-translations

- **Given word-alignments** (produced automatically or otherwise), we do *not* need to do EM training. E.g.,
  - $P(f_1 f_2 | e_1 e_2 e_3) = \frac{Count(f_1 f_2, e_1 e_2 e_3)}{Count(e_1 e_2 e_3)}$



Phrase-based translation in practice




decoding
models

Decoding

- **Decoding** is the act of translating another language into your native language.
  - Decoding is an NP-complete problem (Knight,1999).

- IBM Models often decoded with **stack decoding** or **A* search**.

- **Seminal paper**: U. Germann, M. Jahr, K. Knight, D. Marcu, K. Yamada (2001) *Fast Decoding and Optimal Decoding for Machine Translation.* In: ACL-2001.
  - Introduces **greedy decoding** – start with a solution and incrementally try to **improve** it.

First stage of greedy method

- For each French word f_j, pick the English word e^* such that

$$e^* = \underset{e}{\operatorname{argmax}}\, P(f_j|e)$$

- This gives an initial alignment, e.g.,

Bien	*entendu*	*,*	*il*	*parle*	*d'*	*une*	*belle*	*victoire*
Well	*heard*	*,*	*it*	*talking*	*∅*	*a*	*beautiful*	*victory*

(Better: *quite naturally, he talks about a great victory*)

Some transformations

- $\boldsymbol{Change(j, e)}$: sets translation of $\boldsymbol{f_j}$ to $\boldsymbol{e}$
  - Usually we only consider English words e that are in the top N ranked translations for f_j.

- $\boldsymbol{Change2(j_1, e1, j_2, e2)}$: sets translation of $\boldsymbol{f_{j_1}}$ to $\boldsymbol{e1}$ and translation of $\boldsymbol{f_{j_2}}$ to $\boldsymbol{e2}$
  - Like performing **two** $Change$ transformations in sequence, but **without** evaluating the intermediate string.

- $\boldsymbol{ChangeAndInsert(j, e1, e2)}$: sets translation of $\boldsymbol{f_j}$ to $\boldsymbol{e1}$ and inserts $\boldsymbol{e2}$ at its most likely position.

Some more transformations

- $\boldsymbol{RemoveInfertile(i)}$: Removes $\boldsymbol{e_i}$ if $\boldsymbol{e_i}$ is aligned with *no* French words.
- $\boldsymbol{SwapSeg(i_1, i_2, j_1, j_2)}$: Swaps segment $\boldsymbol{e_{i_1:i_2}}$ with segment $\boldsymbol{e_{j_1:j_2}}$ such that segments do not overlap.
- $\boldsymbol{JoinWords(i_1, i_2)}$: Removes $\boldsymbol{e_{i_1}}$ and *aligns* all French words that were aligned to $\boldsymbol{e_{i_1}}$ to $\boldsymbol{e_{i_2}}$.

Iterating greedily

- We have an initial pair $(E^{(0)}, a^{(0)})$.
- Use local **transformations** to map (E, a) to new pairs, (E', a').
- **At each iteration, k, take the highest probability pair from all possible transformations**
  - i.e., if $\mathcal{R}(\boldsymbol{E}^{(k)}, \boldsymbol{a}^{(k)})$ is the set of all $(\boldsymbol{E}, \boldsymbol{a})$ 'reachable' from $(\boldsymbol{E}^{(k)}, \boldsymbol{a}^{(k)})$, then at each iteration:

$$\left(E^{(k+1)}, a^{(k+1)}\right) = \underset{(E,a)\in\mathcal{R}(\boldsymbol{E}^{(k)},\boldsymbol{a}^{(k)})}{\operatorname{argmax}} P(E)P(F, a|E)$$

UNIVERSITY OF TORONTO

Example of greedy search

Bien	*entendu*	*,*	*il*	*parle*	*d'*	*une*	*belle*	*victoire*
Well	*heard*	*,*	*it*	***talking***	*∅*	*a*	***beautiful***	*victory*

$$\boldsymbol{Change2}(5, talks, 8, great)$$

Bien	*entendu*	*,*	*il*	*parle*	*d'*	*une*	*belle*	*victoire*
Well	*heard*	*,*	*it*	***talks***	*∅*	*a*	***great***	*victory*

Example of greedy search

Bien	entendu	,	il	parle	d’	une	belle	victoire
Well	***heard***	,	*it*	*talks*	$\emptyset$	*a*	*great*	*victory*

$$\boldsymbol{Change2}(2, understood, 6, about)$$

Bien	entendu	,	il	parle	d’	une	belle	victoire
Well	***understood***	,	*it*	*talks*	***about***	*a*	*great*	*victory*

UNIVERSITY OF TORONTO

Example of greedy search

Bien	*entendu*	*,*	*il*	*parle*	*d'*	*une*	*belle*	*victoire*
Well	*understood*	*,*	***it***	*talks*	*about*	*a*	*great*	*victory*

$$\boldsymbol{Change}(4, he)$$

Bien	*entendu*	*,*	*il*	*parle*	*d'*	*une*	*belle*	*victoire*
Well	*understood*	*,*	***he***	*talks*	*about*	*a*	*great*	*victory*

UNIVERSITY OF TORONTO

Example of greedy search

Bien	*entendu*	*,*	*il*	*parle*	*d'*	*une*	*belle*	*victoire*
Well	***understood***	*,*	*he*	*talks*	*about*	*a*	*great*	*victory*

$$\boldsymbol{Change2}(1, quite, 2, naturally)$$

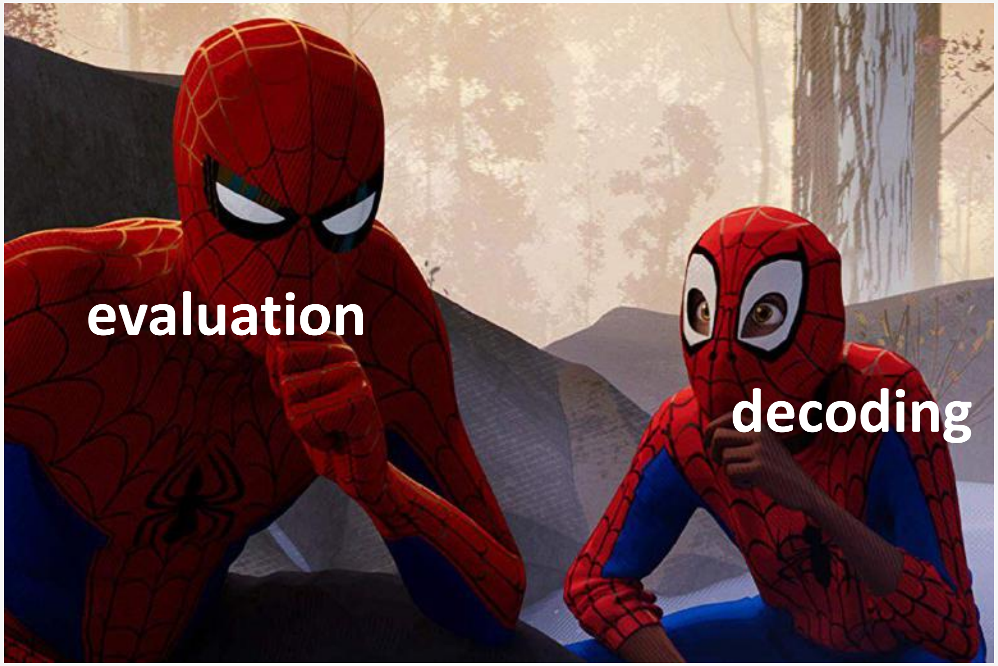
Bien	*entendu*	*,*	*il*	*parle*	*d'*	*une*	*belle*	*victoire*
Quite	***naturally***	*,*	*he*	*talks*	*about*	*a*	*great*	*victory*

Greedy transformations

- At each iteration, we try *each possible* transformation.
- For each possible transformation, we evaluate

$$P(E)P(F, a|E)$$

- We choose the transformation that gives the highest probability, and iterate until some stopping condition.


evaluation
decoding

Evaluation of MT systems

对外经济贸易合作部今天提供的数据表明，今年至十一月中国实际利用外资四百六十九点五九亿美元，其中包括外商直接投资四百点零七亿美元。

Human	According to the data provided today by the Ministry of Foreign Trade and Economic Cooperation, as of November this year, China has actually utilized 46.959B US dollars of foreign capital, including 40.007B US dollars of direct investment from foreign businessmen.
IBM4	The Ministry of Foreign Trade and Economic Cooperation, including foreign direct investment 40.007B US dollars today provide data include that year to November China actually using foreign 46.959B US dollars and
Yamada/ Knight	Today’s available data of the Ministry of Foreign Trade and Economic Cooperation shows that China’s actual utilization of November this year will include 40.007B US dollars for the foreign direct investment among 46.959B US dollars in foreign capital.

How can we objectively compare the quality of two translations?

Automatic evaluation

- We want an **automatic** and effective method to **objectively** rank competing translations.
  - **Word Error Rate (WER)** measures the number of erroneous word **insertions**, **deletions**, **substitutions** in a translation.
    - E.g., **Reference**: *how to recognize speech*
      Translation: *how understand a speech*

    - **Problem**: There are many possible valid translations.
      (There's no need for an exact match)

UNIVERSITY OF TORONTO

Challenges of evaluation

- **Human judges**: expensive, slow, non-reproducible (different judges – different biases).

- Multiple valid translations, e.g.:
  - **Source**: *Il s'agit d'un guide qui assure que l'armée sera toujours fidèle au Parti*
  - **T1**: *It is a guide to action that ensures that the military will forever heed Party commands*
  - **T2**: *It is the guiding principle which guarantees the military forces always being under command of the Party*

UNIVERSITY OF TORONTO

BLEU evaluation

- **BLEU (BiLingual Evaluation Understudy)** is an automatic and popular method for evaluating MT.
  - It uses **multiple** human **reference** translations, and looks for local matches, allowing for phrase movement.

  - **Candidate:** *n.* a translation produced by a machine.

- There are a few parts to a **BLEU score**…

Example of BLEU evaluation

- **Reference 1**: *It is a guide to action that ensures that the military will forever heed Party commands*
- **Reference 2**: *It is the guiding principle which guarantees the military forces always being under command of the Party*
- **Reference 3**: *It is the practical guide for the army always to heed the directions of the party*

- **Candidate 1**: *It is a guide to action which ensures that the military always obeys the commands of the party*
- **Candidate 2**: *It is to insure the troops forever hearing the activity guidebook that party direct*

BLEU: Unigram precision

- The **unigram precision** of a candidate is

$$\frac{C}{N}$$

where N is the number of words in the **candidate**
and C is the number of words in the **candidate** which are in **at least one reference**.

- e.g., **Candidate 1**: *It is a guide to action which ensures that the military always obeys the commands of the party*
  - **Unigram precision** $= \frac{17}{18}$
    (*obeys* appears in none of the three references).

BLEU: Modified unigram precision

- **Reference 1**: *The lunatic is on the grass*
- **Reference 2**: *There is a lunatic upon the grass*
- **Candidate**: *The the the the the the the*
  - Unigram precision $= \frac{7}{7} = 1$

- **Capped unigram precision:**

  A candidate word type w can only be correct a **maximum** of $cap(w)$ times.
  - e.g., with $\boldsymbol{cap(the) = 2}$, the above gives

    $$p_1 = \frac{2}{7}$$

BLEU: Generalizing to *N*-grams

- Generalizes to higher-order *N*-grams.
  - **Reference 1**: *It is a guide to action that ensures that the military will forever heed Party commands*
  - **Reference 2**: *It is the guiding principle which guarantees the military forces always being under command of the Party*
  - **Reference 3**: *It is the practical guide for the army always to heed the directions of the party*

  - **Candidate 1**: *It is a guide to action which ensures that the military always obeys the commands of the party*
  - **Candidate 2**: *It is to insure the troops forever hearing the activity guidebook that party direct*

Bigram precision, p_2

$p_2 = 10/17$

$p_2 = 1/13$

BLEU: Precision is not enough

- **Reference 1**: *It is a guide to action that ensures that the military will forever heed Party commands*
- **Reference 2**: *It is the guiding principle which guarantees the military forces always being under command **of the** Party*
- **Reference 3**: *It is the practical guide for the army always to heed the directions **of the** party*

- **Candidate 1**: ***of the***

Unigram precision, $p_1 = \frac{2}{2} = 1$ Bigram precision, $p_2 = \frac{1}{1} = 1$

BLEU: Brevity

- Solution: Penalize brevity.
- **Step 1**: for each candidate,
  find the reference **most similar in length**.
- **Step 2**: $\boldsymbol{c_i}$ is the length of the i^{th} candidate, and
  $\boldsymbol{r_i}$ is the nearest length among the references,

$$brevity_i = \frac{r_i}{c_i}$$

Bigger = too brief

- **Step 3**: multiply precision by the (0..1) **brevity penalty**:

$$BP = \begin{cases} 1 & \text{if } brevity < 1 \\ e^{1-brevity} & \text{if } brevity \geq 1 \end{cases}$$

$(r_i < c_i)$

$(r_i \geq c_i)$

BLEU: Final score

- On slide 39, $r_1 = 16, r_2 = 17, r_3 = 16$, and $c_1 = 18$ and $c_2 = 14$,

$$brevity_1 = \frac{17}{18} \qquad BP_1 = 1$$

$$brevity_2 = \frac{16}{14} \qquad BP_2 = e^{1-\left(\frac{8}{7}\right)} = 0.8669$$

- **Final score** of candidate C:

$$\boxed{BLEU = BP_C \times (p_1 p_2 \dots p_n)^{1/n}}$$

where p_n is the n-gram precision. (You can set n empirically)

Example: Final BLEU score

- **Reference 1:** *I am afraid Dave*
  Reference 2: *I am scared Dave*
  Reference 3: *I have fear David*
  Candidate: *I fear David*

Assume $cap(\cdot) =$ 2 for all *N*-grams

Also assume BLEU order $n = 2$

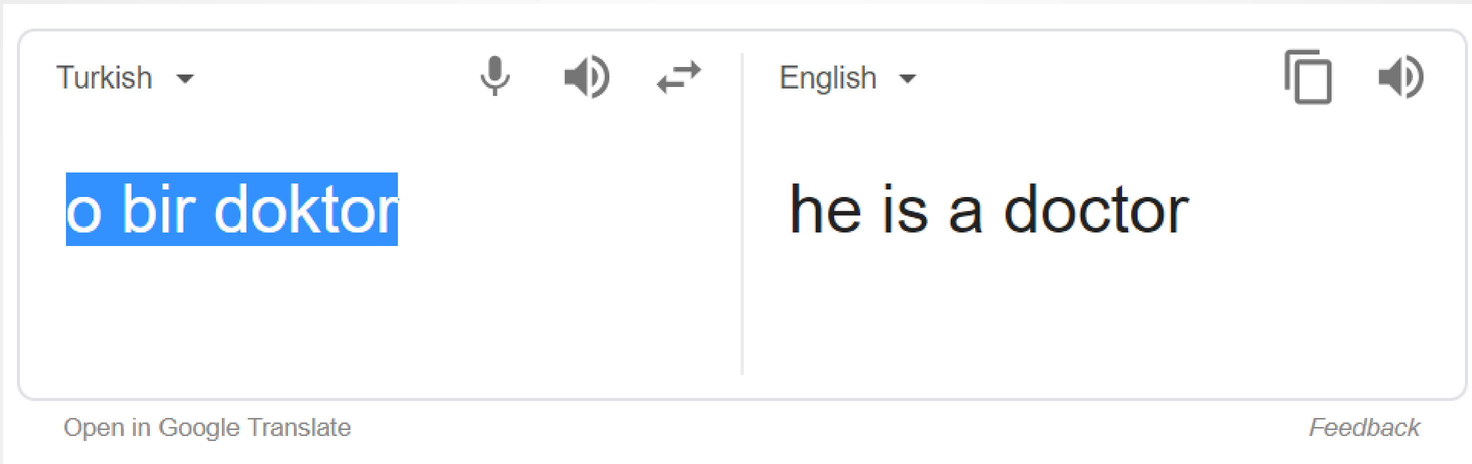
- $brevity = \frac{4}{3} \geq 1$ so $BP = e^{1-\left(\frac{4}{3}\right)}$
- $p_1 = \frac{1+1+1}{3} = 1$
- $p_2 = \frac{1}{2}$
- $BLEU = BP(p_1 p_2)^{\frac{1}{2}} = e^{1-\left(\frac{4}{3}\right)} \left(\frac{1}{2}\right)^{\frac{1}{2}} \approx 0.5067$

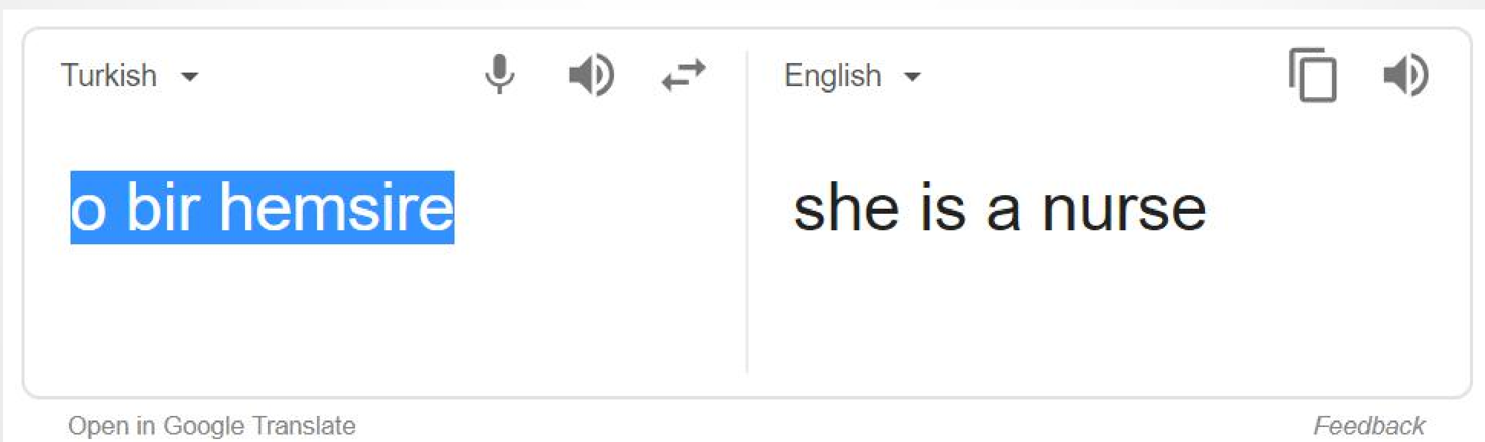
BLEU: summary

- BLEU is a geometric mean over n-gram precisions.
  - These precisions are **capped** to avoid strange cases.
    - E.g., the translation "*the the the the*" is not favoured.
  - This geometric mean is **weighted** so as not to favour unrealistically short translations, e.g., "*the*"
- Initially, evaluations showed that BLEU predicted human judgements very well, but:
  - People started **optimizing** MT systems to **maximize** BLEU. Correlations between BLEU and humans **decreased**.

UNIVERSITY OF TORONTO

(Aside) Bias in machine translation

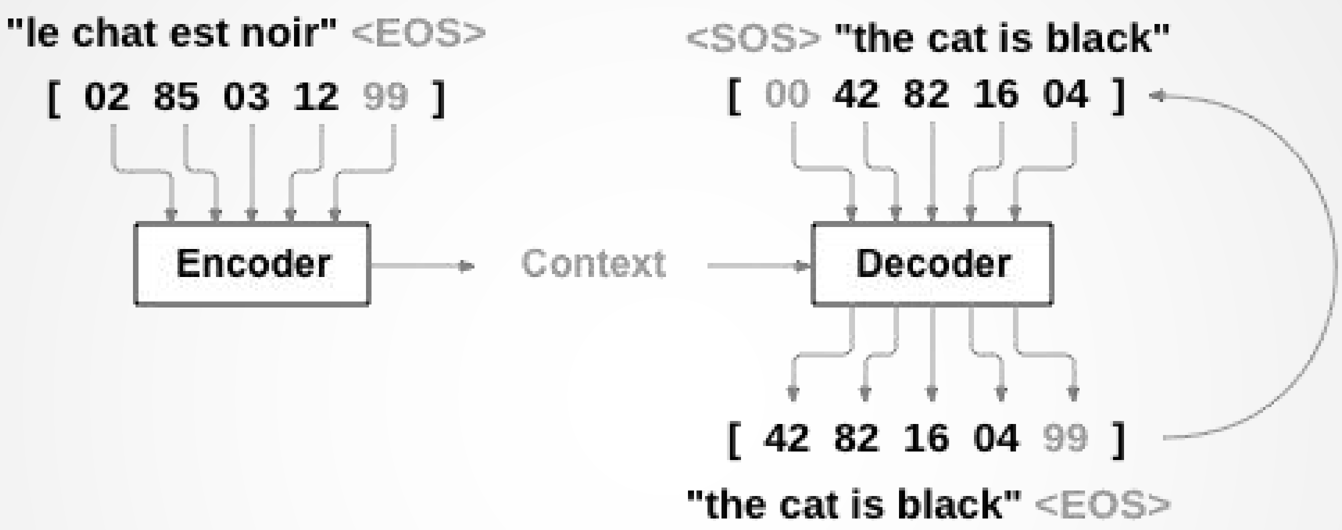




(Aside) Other evaluation methods

- **METEOR** is a weighted F-measure (combination of recall and precision)
- **Translation Error Rate** computes the string edit distance between the reference and the hypothesis.
- The **RIBES** metric looks at rank correlation to measure word order similarity between system and reference translations.

(Preview) Neural machine translation



Reading

- **Entirely optional**: Vogel, S., Ney, H., and Tillman, C. (1996). *HMM-based Word Alignment in Statistical Translation*. In: Proceedings of the 16th International Conference on Computational Linguistics, pp. 836-841, Copenhagen.

- (optional) Gale & Church "*A Program for Aligning Sentences in Bilingual Corpora*" (on course website)

- **Useful reading on IBM Model-1**: Section 25.5 of the 2nd edition of the Jurafsky & Martin text.
  - 1st edition available at Robarts library.

- **Other**: Manning & Schütze Sections 13.0, 13.1.2 (Gale&Church), 13.1.3 (Church), 13.2, 13.3, 14.2.2