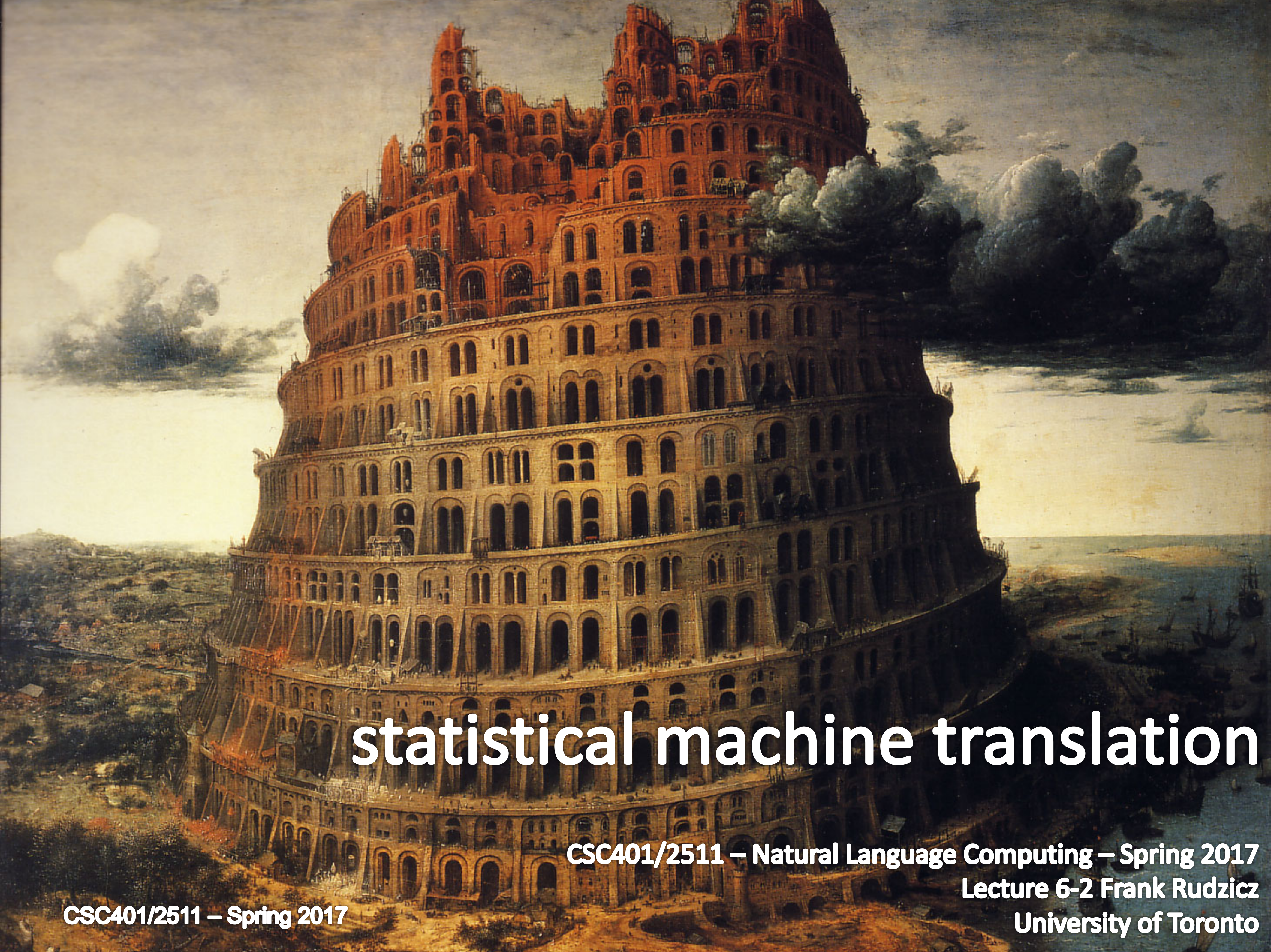


statistical machine translation

CSC401/2511 – Natural Language Computing – Spring 2017
Lecture 6-2 Frank Rudzicz
University of Toronto

Today

- Overview of IBM-2, IBM-3, and phrase-based methods.
- Decoding for SMT.
- Evaluation of MT systems.

Practical note on programming IBM-1

- If you were to code the EM algorithm for IBM-1, you would **not** initialize $\theta = P(f|e)$ uniformly over the **entire** vocabulary.
  - Don't make a $V_F \times V_E$ table with $P(f|e) = 1/\|V_E\|$ X

- This structure would be too **large**.
  - Probabilities would be too **small**.
  - It would take **too much work** to update.

- Rather, initialize a **hash table** over ***possible*** alignments, $\mathcal{M}$. For every English word e, only consider French words f in sentences **aligned** with English sentences containing e.
  - e.g., structure $\mathrm{P}.e.f := P(f|e) = 1/\|\mathcal{M}\|$

UNIVERSITY OF TORONTO

Higher IBM models

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

- Only IBM Model 1 training reaches a *global maximum*
  - Training of each IBM model **extends** the next **next lowest** model.

- **Higher models** become computationally **expensive**.

UNIVERSITY OF TORONTO

IBM-2

- **Unlike IBM Model-1**, the placement of a word in, say, Spanish in **IBM Model-2** depends on where its **equivalent** word was in English.
  - IBM-2 captures the intuition that translations should lie roughly “along the diagonal”.

	Buenos	dias	,	me	gusta	papas	frías
Good	X						
day		X					
,			X				
I				X			
like					X		
cold							X
potatoes						X	

UNIVERSITY OF TORONTO

IBM-2

- IBM Model 2 **builds** on Model 1 by adding a **re-ordering model** defined by **distortion** parameters ***<u>regardless of actual words</u>***.

$D(i|j,\mathcal{L}_E,\mathcal{L}_F)$= the probability that the i^{th} English **slot** is aligned to the j^{th} French **slot**, given sentence lengths $\mathcal{L}_E$ and $\mathcal{L}_F$.

- In IBM Model 2:

$$P(a|E,\mathcal{L}_E,\mathcal{L}_F) = \prod_{j=1}^{\mathcal{L}_F} D(a_j|j,\mathcal{L}_E,\mathcal{L}_F)$$

- Recall that in IBM Model 1,

$$P(a|E,\mathcal{L}_E,\mathcal{L}_F) = \frac{P(\mathcal{L}_F)}{(\mathcal{L}_E+1)^{\mathcal{L}_F}}$$

IBM-2 – Probability of alignment

- E = *And the program has been implemented*
- F = *Le programme a été mis en application*
- $\mathcal{L}_E = 6$
- $\mathcal{L}_F = 7$
- $a = \{2,3,4,5,6,6,6\}$ (i.e., $f_1 \leftarrow e_2$, $f_2 \leftarrow e_3$,...)

D(2nd English word | 1st French word,...)

- $P(a|E, \mathcal{L}_E, \mathcal{L}_F) = D(2|1,6,7) \times D(3|2,6,7) \times D(4|3,6,7) \times D(5|4,6,7) \times D(6|5,6,7) \times D(6|6,6,7) \times D(6|7,6,7)$

This is independent of the actual **words**.
This cares only about **position**.

UNIVERSITY OF TORONTO

IBM-2: generation

- To **generate** a French sentence F from English E,
  1. **Pick an alignment** with probability
     $$\prod_{j=1}^{\mathcal{L}_F} D(a_j|j, \mathcal{L}_E, \mathcal{L}_F)$$
  3. **Sample** French words with probability
     $$P(F|a, E) = \prod_{j=1}^{\mathcal{L}_F} P(f_j|e_{a_j})$$

This is the same $P(f|e)$ as in IBM-1.

So,

$$P(F, a|E) = P(a|E)P(F|a, E) = \prod_{j=1}^{\mathcal{L}_F} D(a_j|j, \mathcal{L}_E, \mathcal{L}_F)P(f_j|e_{a_j})$$

UNIVERSITY OF TORONTO

IBM-2: training

- We use EM, as before with IBM-1 **except** that we need to take the **distortion** into account when computing the probability of an alignment.

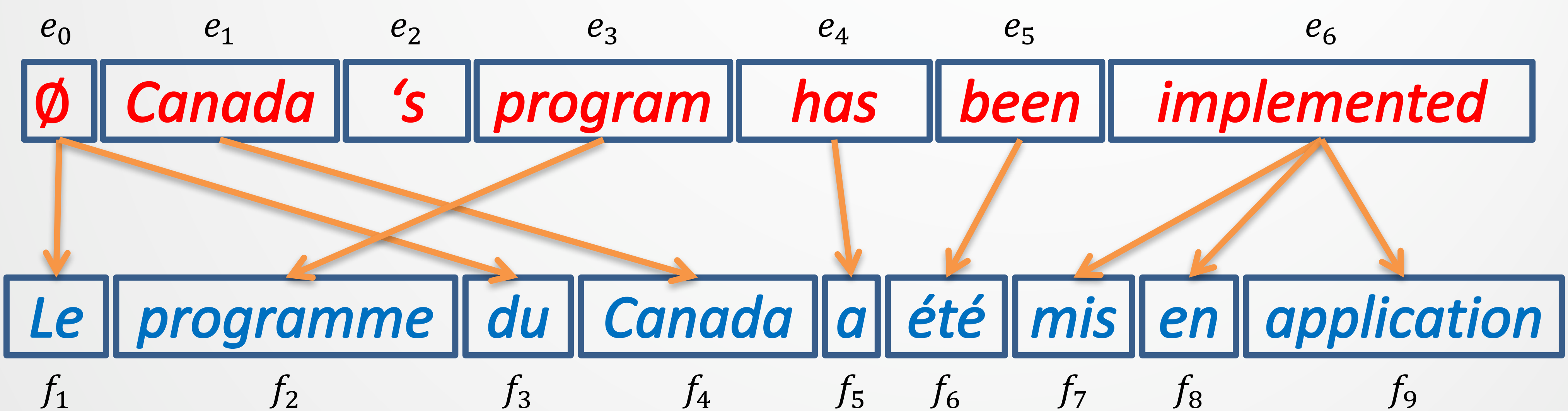
- We also need to **learn** the distortion function.

- *Aren't you glad that you don't need to know how to compute EM for IBM-2?*

UNIVERSITY OF TORONTO

IBM-3

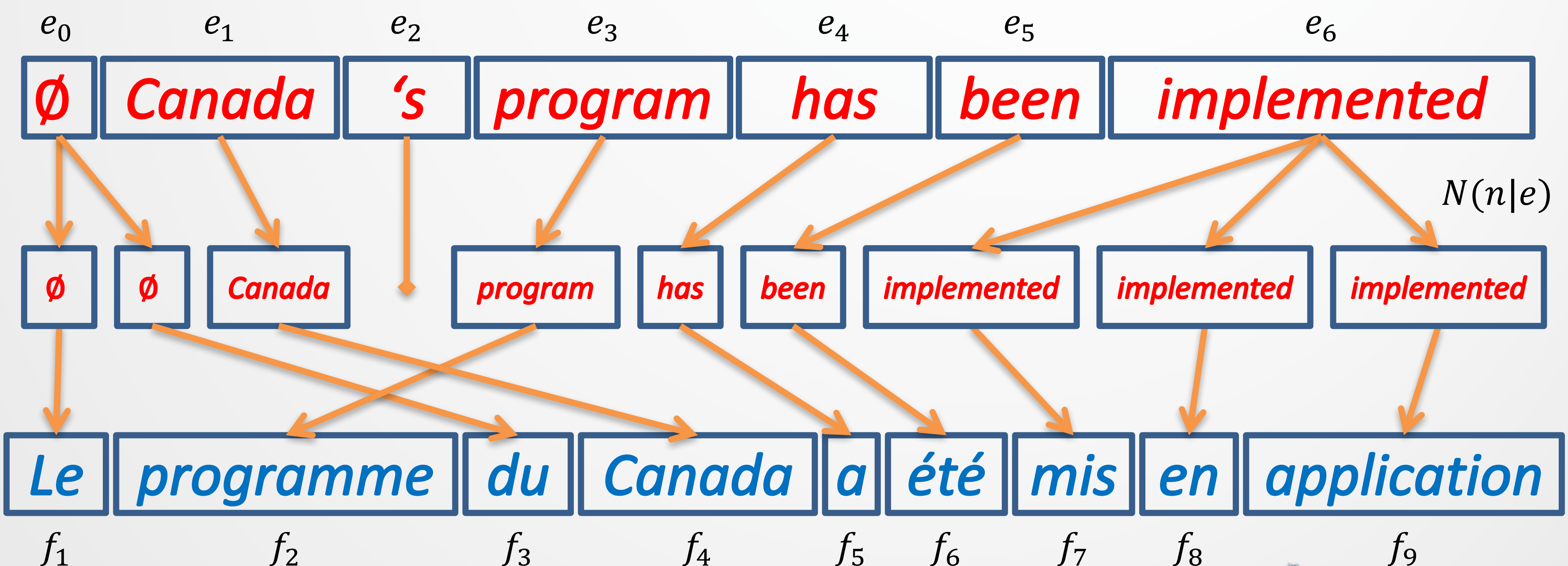
- **IBM Model 3** extends Model 2 by adding a **fertility model** that describes how many French words each English word can produce.
  - In the example below, *implemented* appears to be **more fertile** than *program*.



UNIVERSITY OF TORONTO

IBM-3: The generation model

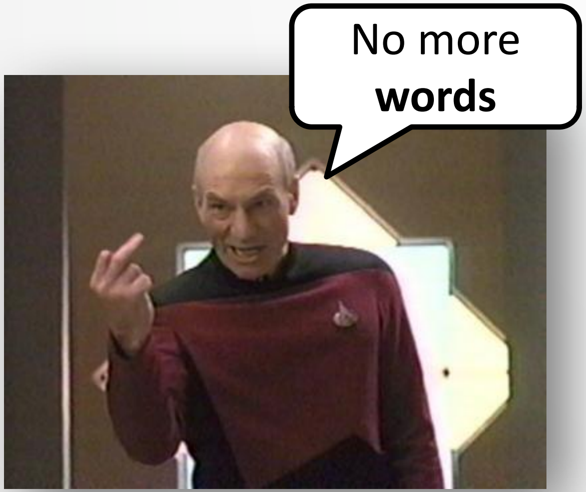
- First, we **replicate** each word according to a new hidden parameter, $\boldsymbol{N(n|e)}$, which is the **probability that word *e* produces *n* words**.
  - We then **re-align** (*with distortion*) and **translate** as we did in IBM-2.



UNIVERSITY OF TORONTO

IBM models

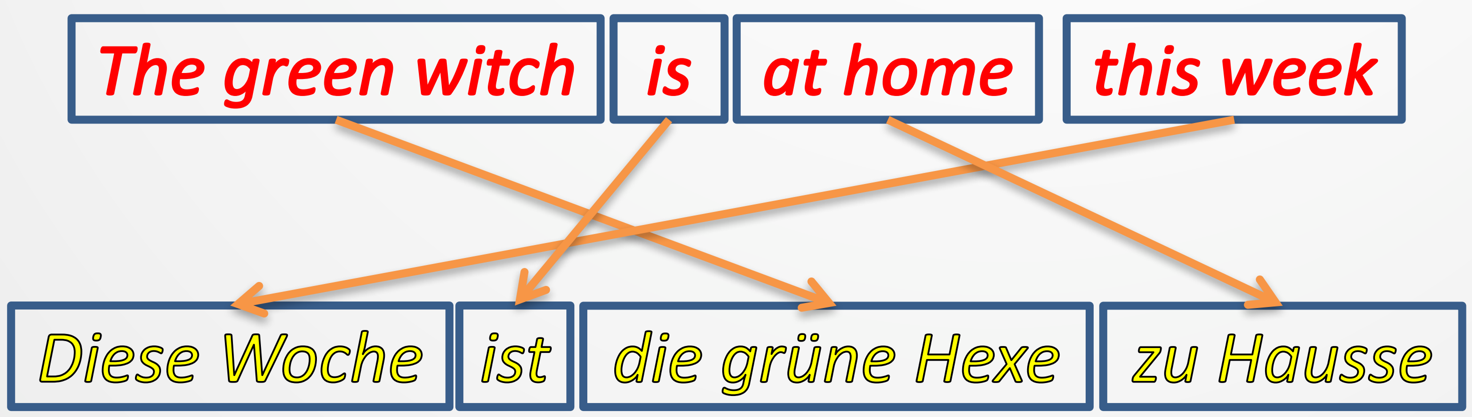
IBM Model 1	lexical translation
IBM Model 2	adds absolute **re-ordering model**
IBM Model 3	adds **fertility model**


No more
words


I want to see
phrases

Phrase-based statistical MT

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



UNIVERSITY OF TORONTO

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.

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

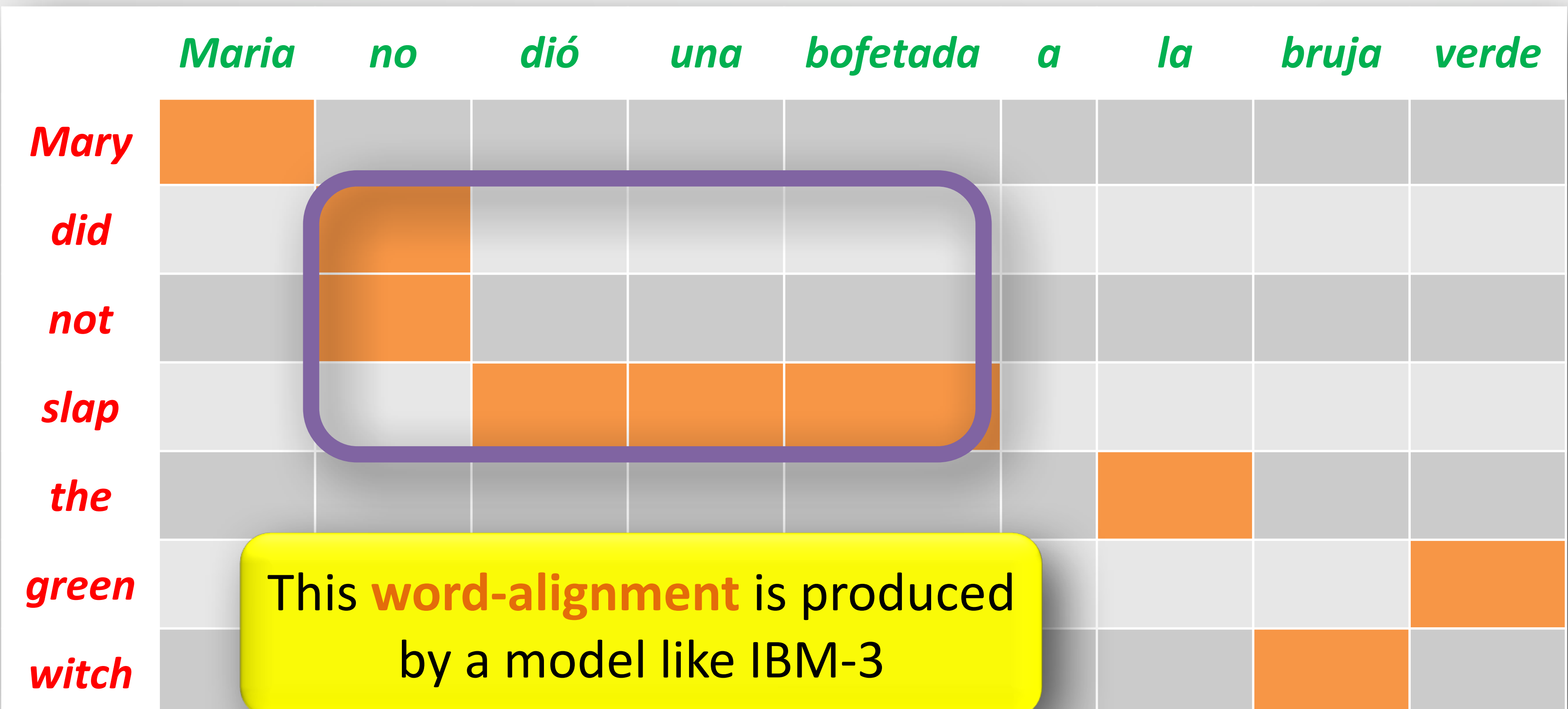
$$P(coup\ d'oeil|glance) > P(coup\ d'oeil|stroke\ of\ eye)$$

No context ☹

A tiny amount of context ☺

Learning phrase-translations

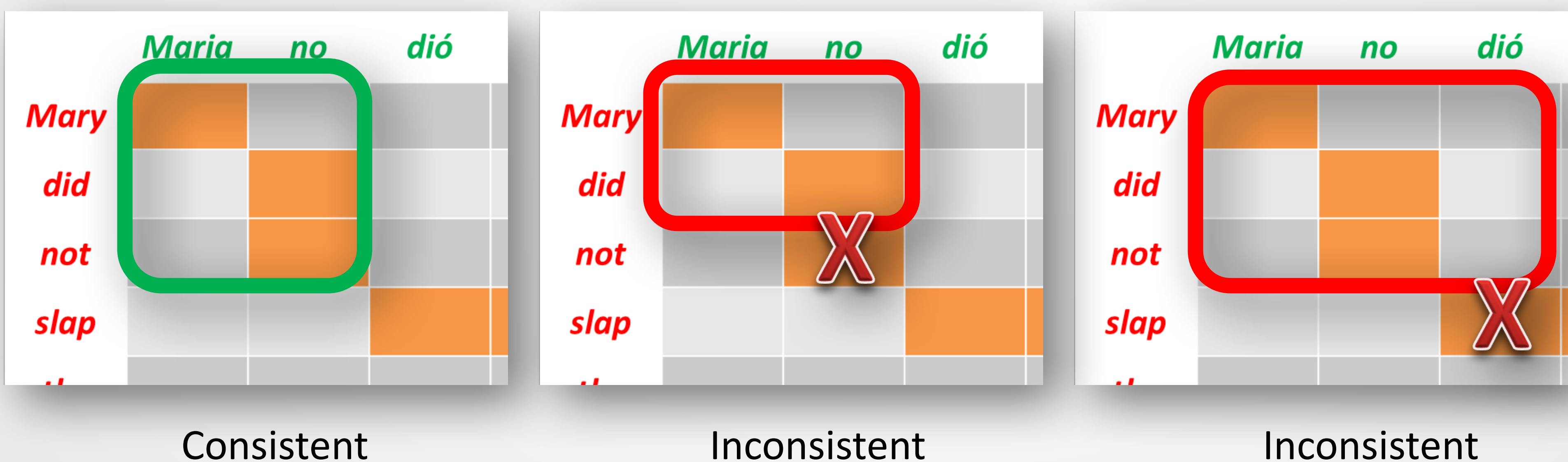
- Typically, we use **alignment templates** (Och *et al.*, 1999).
  - Start with a **word-alignment**, then build **phrases**.



UNIVERSITY OF TORONTO

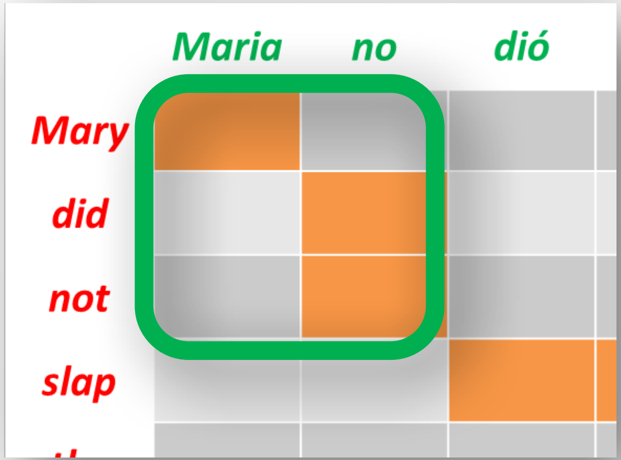
Learning phrase-translations

- 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.



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


No more **models**


I want to see **decoding**

Decoding

- **Decoding** is the act of translating a 'foreign' 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.

UNIVERSITY OF TORONTO

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*)

UNIVERSITY OF TORONTO

Some transformations

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

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

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

UNIVERSITY OF TORONTO

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 $\boldsymbol{e_{i_2}}$.

UNIVERSITY OF TORONTO

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

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

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

Example of greedy search

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

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

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

Example of greedy search

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

$Change(4, he)$

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

Example of greedy search

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

$$Change2(1, quite, 2, naturally)$$

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

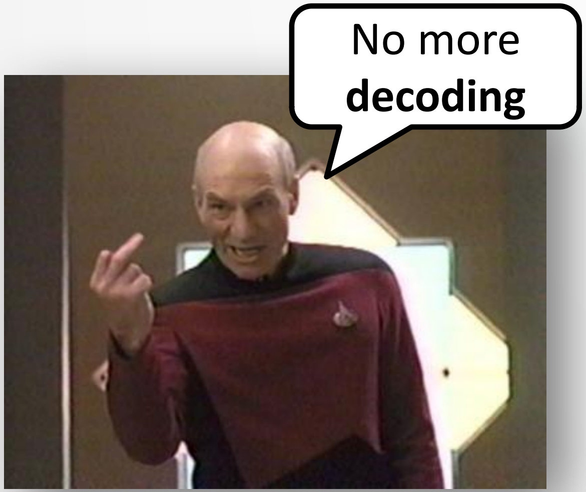
UNIVERSITY OF TORONTO

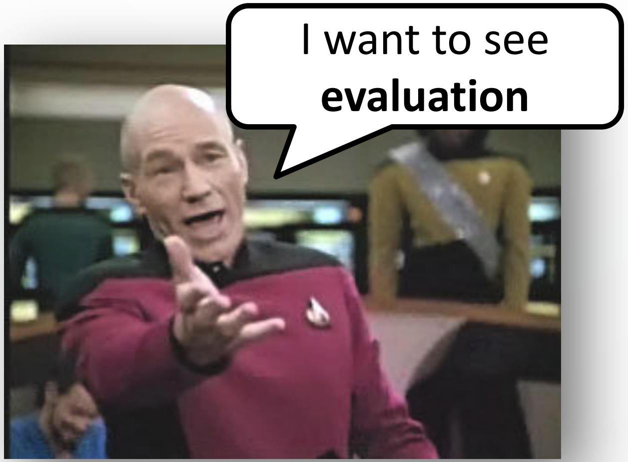
Greedy transformations

- 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.


No more **decoding**


I want to see **evaluation**

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?

UNIVERSITY OF TORONTO

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*

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**…

UNIVERSITY OF TORONTO

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

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

UNIVERSITY OF TORONTO

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, $p_2 = \frac{1}{1} = 1$

UNIVERSITY OF TORONTO

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**:

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

($r_i < c_i$)

($r_i \geq c_i$)

UNIVERSITY OF TORONTO

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:

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

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

UNIVERSITY OF TORONTO

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

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.

- **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 13.1.2 (Gale&Church), 13.1.3 (Church), 13.3, 14.2.2

Announcements

- **Assignment 1** marks/comments will be emailed individually.
- **Not-for-marks midterm** on Monday 6 March.

UNIVERSITY OF TORONTO