Lecture 07:
Machine Translation

OVERVIEW

1. Introduction to machine translation
2. Statistical machine translation
3. Difficulties

MACHINE TRANSLATION (MT)

Machine Translation (MT) is the task of translating a sentence *x* from one language (the source language) to another sentence *y* in another language (the target language).

x: L'homme est né libre, et partout il est dans les fers

↓

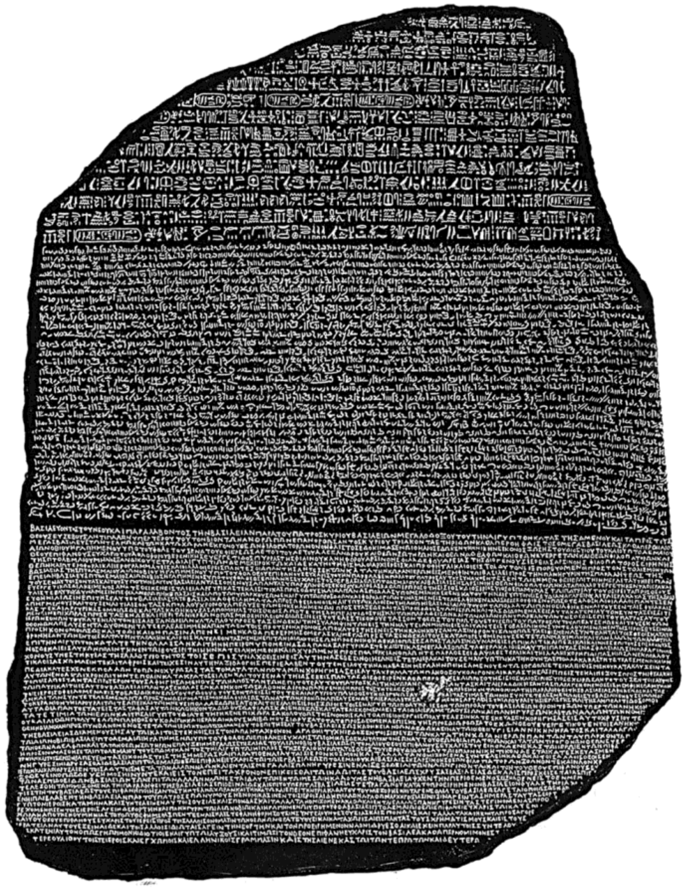
y: Man is born free, but everywhere he is in chains

- Rousseau

The Rosetta Stone

First known historical evidence of translation

Instance of parallel text: Greek inscription allowed scholars to decipher the hieroglyphs



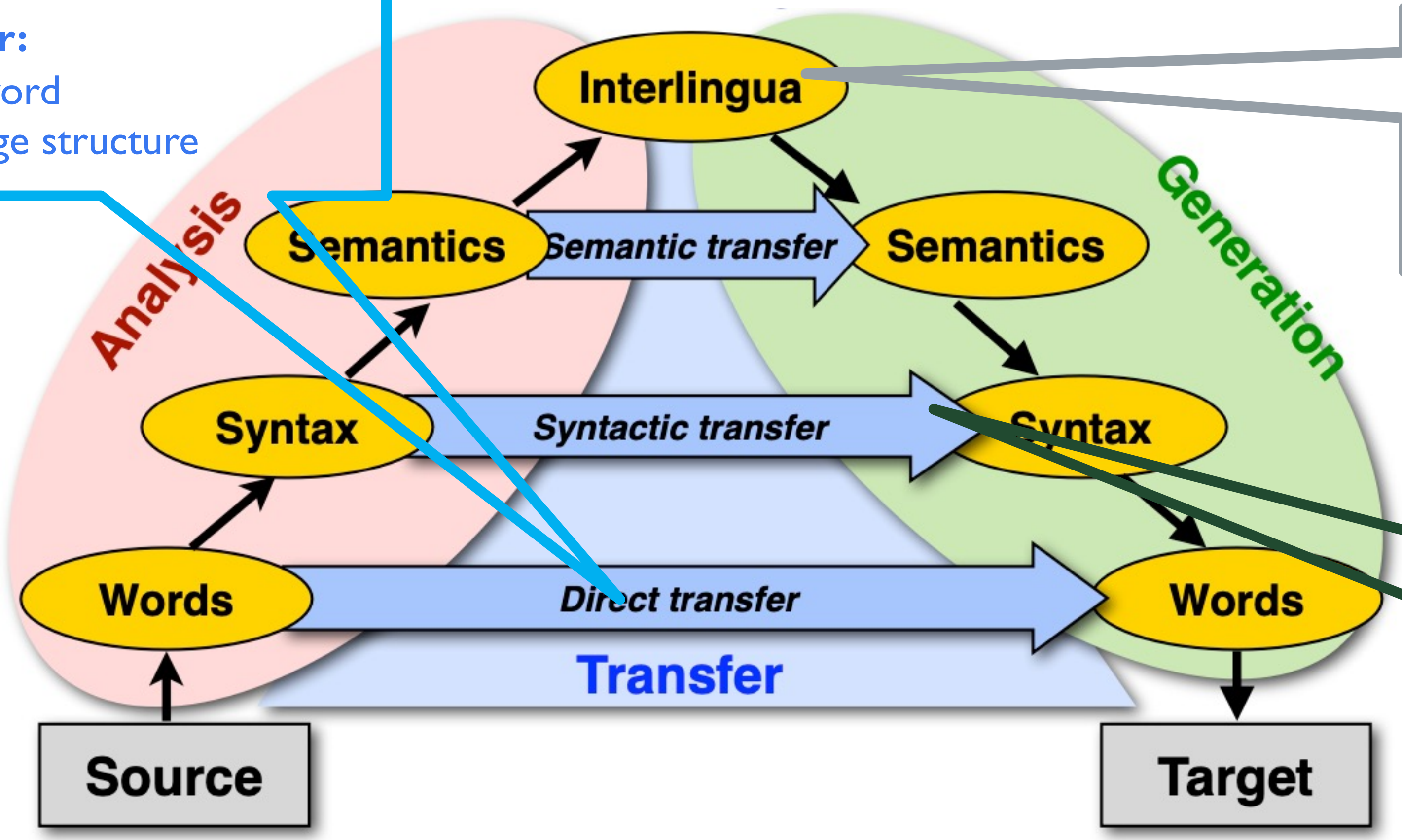
Hieroglyphic: used by priest in ancient Egypt

Demotic: used for daily purposes in Egypt

Ancient Greek: used by the administration

THE VAUQUOIS TRIANGLE

Direct transfer:
- word by word
- No language structure



Interlingual:
- Analyze source language and represent as interlingual
- Generate target from interlingual

Transfer-based:
- Parse source language
- Determine its structure
- apply rules to transfer structure to target language

STATISTICAL MACHINE TRANSLATION (SMT)

- Suppose we want to translate a text from *French* to *English*
- We need to find the *best English sentence* y, given a *French sentence* x $\quad P(y|x), \forall y \in \Omega$

Bayes Rule

$$\underset{y}{\operatorname{argmax}}\, P(y|x) \;=\; \underset{y}{\operatorname{argmax}}\, \underbrace{P(x|y)}\,\underbrace{P(y)}$$

Translation Model

Models how words and phrases should be translated (*fidelity*). Learnt from parallel data.

Language Model

Models how to write good English (*fluency*). Learnt from monolingual data.

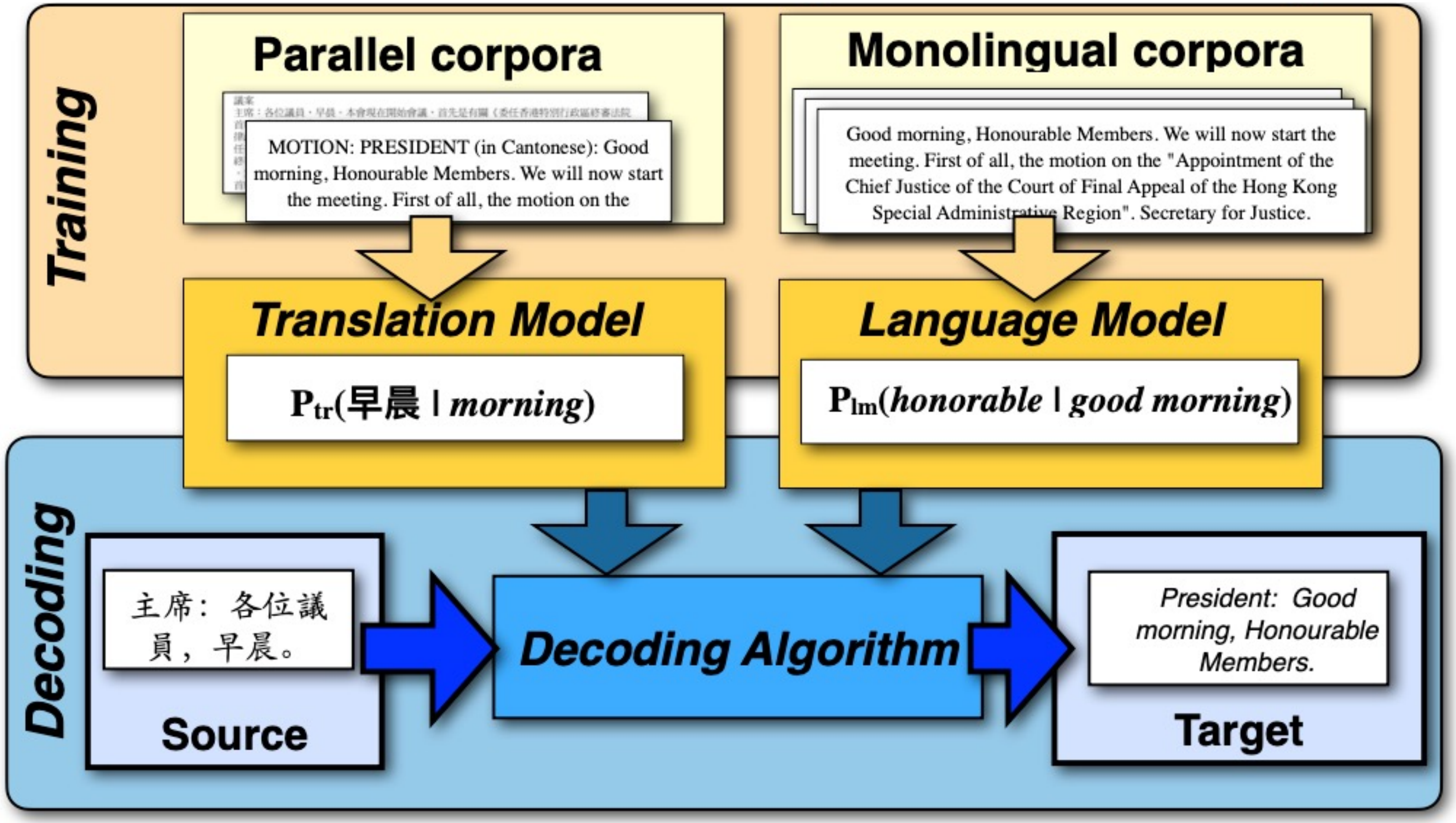
LEARNING ALIGNMENT FOR SMT

$$\underbrace{\text{argmax}_y}\underbrace{P(x|y)}_{\text{Translation Model}}\underbrace{P(y)}_{\text{Language Model}}$$

Question:
How to compute this argmax?

- Enumerate every possible *y* and calculate the probability?
  - Too expensive!
- **Solution:** Use a heuristic search algorithm to search for the best translation, discarding hypotheses with very low-probability
  - This process is called *decoding*

SMT training and decoding



STATISTICAL MACHINE TRANSLATION (SMT)

How do we learn the translation model $P(x|\,y)$?

- large corpus of parallel text (French/English)
- Rewrite the translation model

$$P(x|y) \approx P(x, a|y)$$

where a is an alignment or correspondence

- an alignment is a correspondence between target (French) sentence x and source (English) sentence y
- The alignment can be regarded as the decoder

DECODING IN SMT

- Exhaustive search decoding
  - Find translation that maximizes $P(y \mid x)$
  - Try computing all possible sequences y (too expensive)
    - At each time step we are tracking V possible partial translations

- Beam search decoding
  - On each step of decoder keep track of the k most probable partial translation (hypothesis). K is the beam size
  - Beam search is not guaranteed to find optimal solution
  - More efficient than exhaustive search!

STATISTICAL MACHINE TRANSLATION

We learn the alignment $P(x, a|\, y)$ as a combination of many factors including

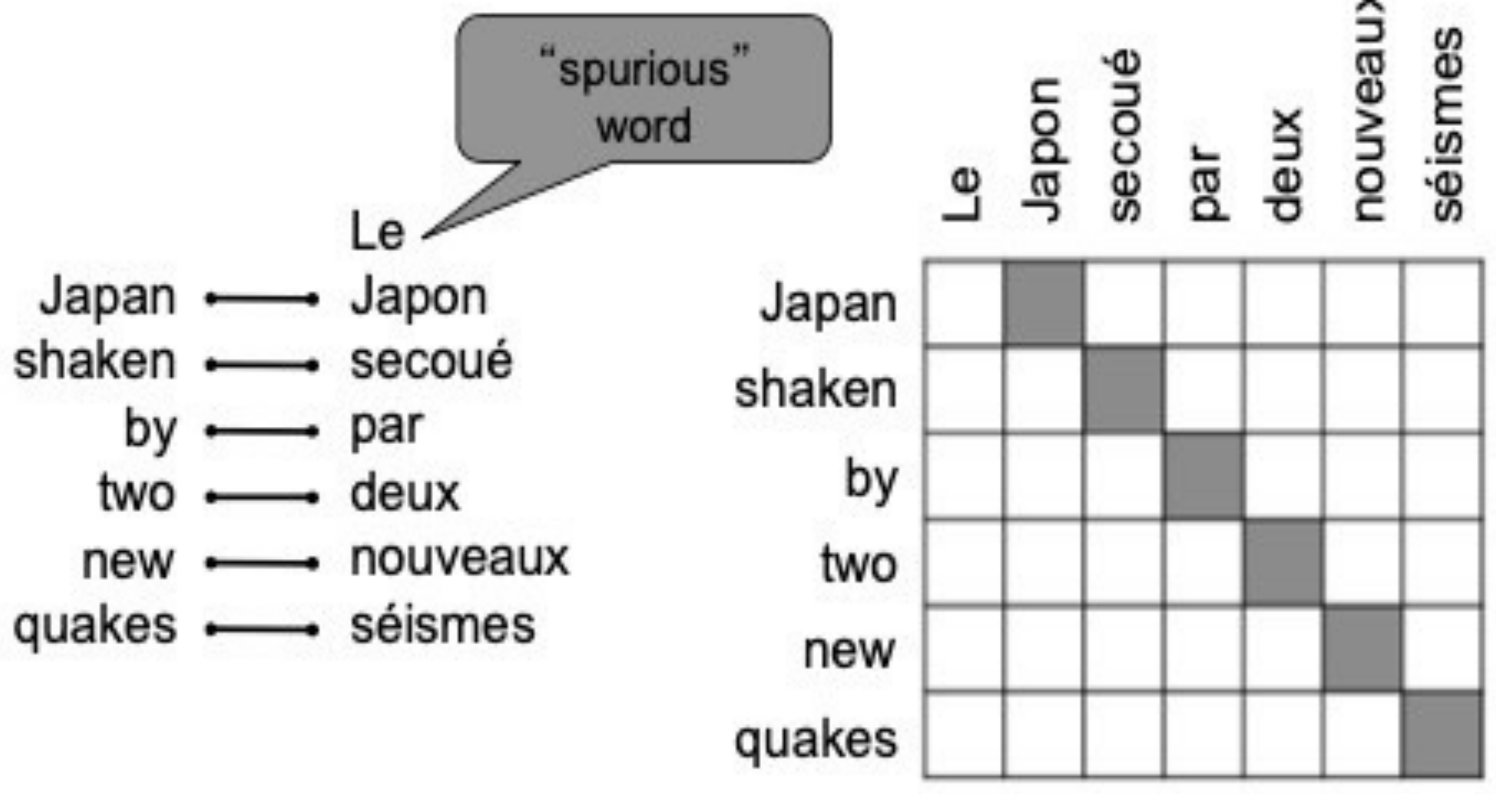
- Probability of particular words aligning
  - can depend on position in sentence

- Probability of particular words having specific fertility
  - One word have correspondence with many words
    - What's the probability of a French word having 3 corresponding English word

NB: Obtaining and alignment decoder in SMT is not trivial task

WHY IS MACHINE TRANSLATION HARD?

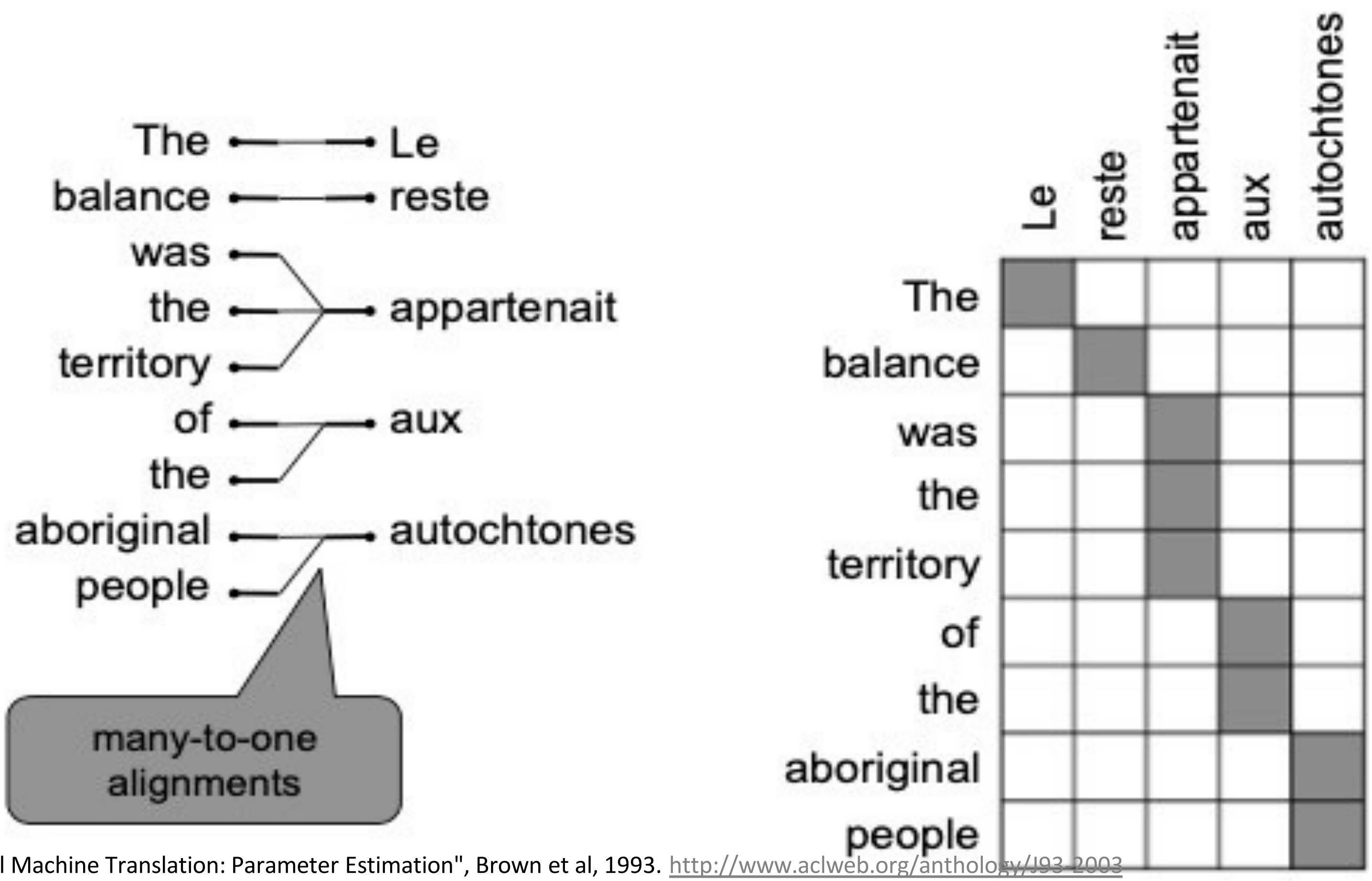
Some words have no counterpart



Examples from: "The Mathematics of Statistical Machine Translation: Parameter Estimation", Brown et al, 1993. http://www.aclweb.org/anthology/J93-2003

WHY IS MACHINE TRANSLATION HARD?

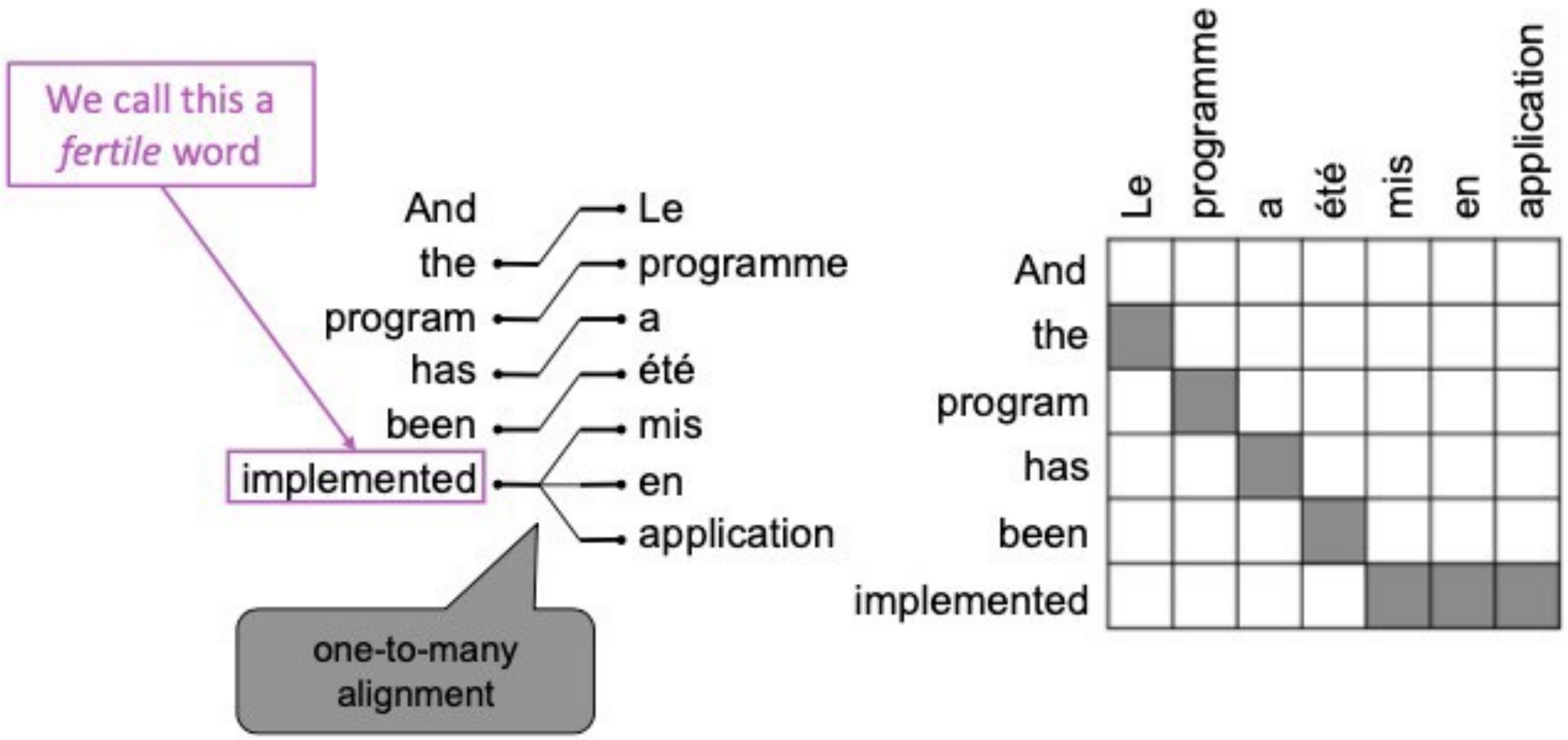
Alignment can be many-to-one



Examples from: “The Mathematics of Statistical Machine Translation: Parameter Estimation", Brown et al, 1993. http://www.aclweb.org/anthology/J93-2003

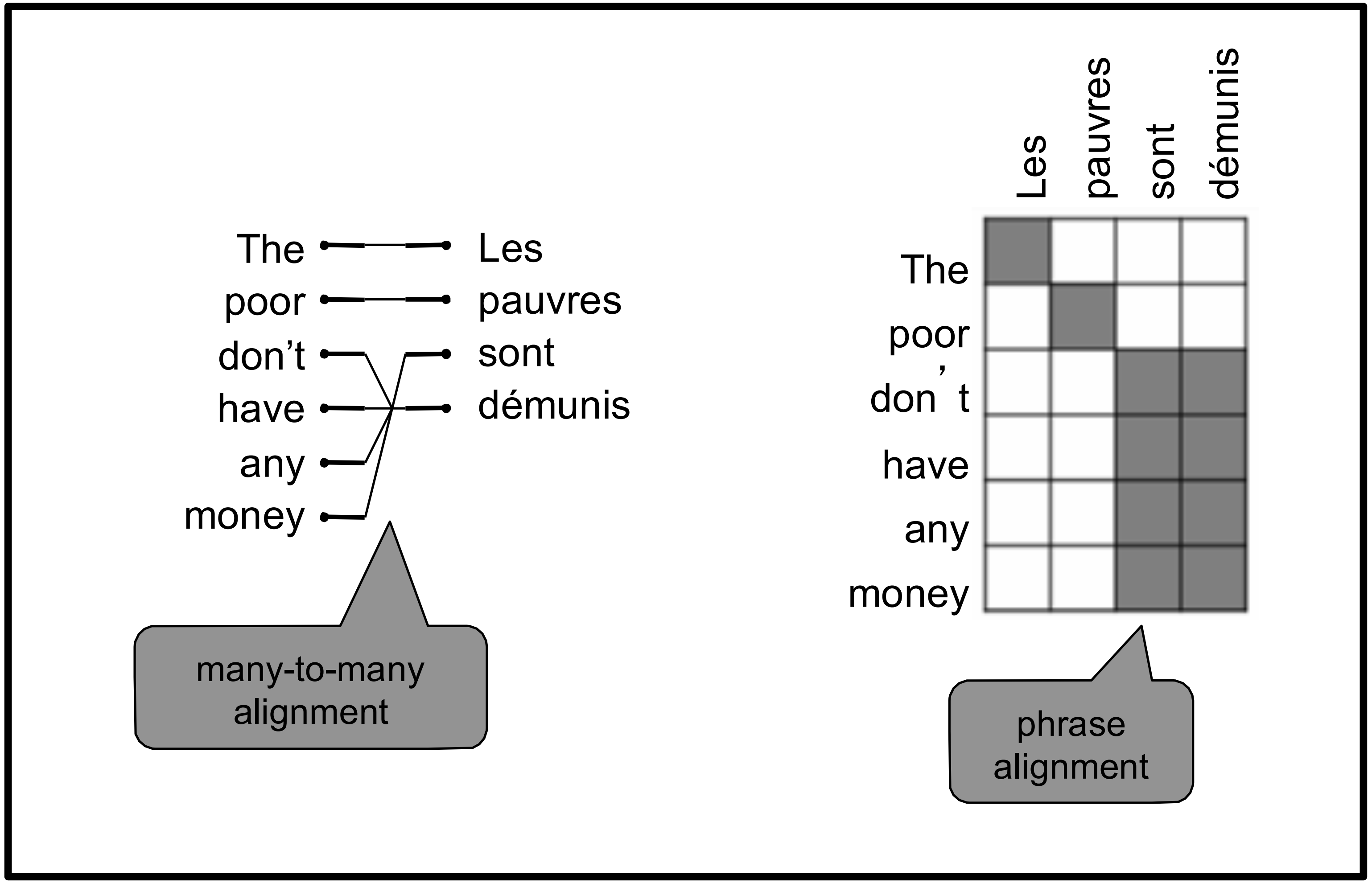
WHY IS MACHINE TRANSLATION HARD?

Alignment can be one-to-many



Examples from: "The Mathematics of Statistical Machine Translation: Parameter Estimation", Brown et al, 1993. http://www.aclweb.org/anthology/J93-2003

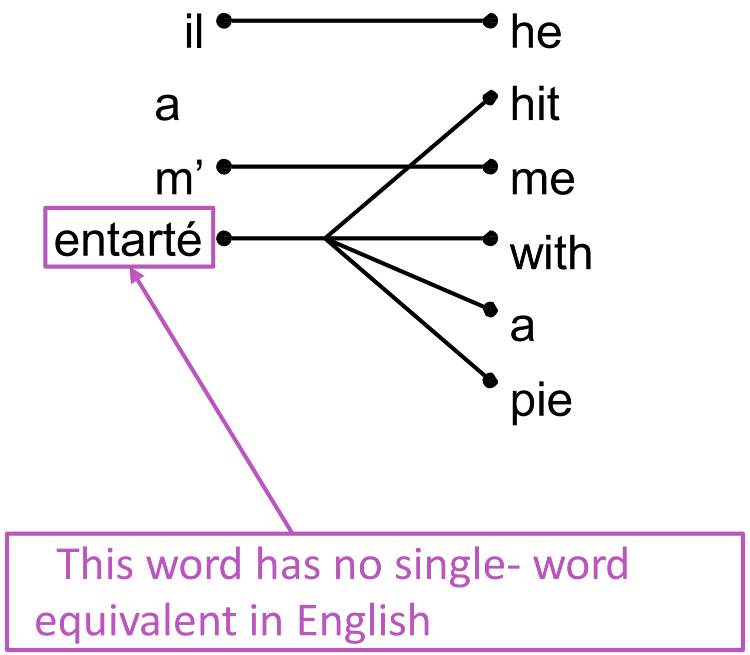
WHY IS MACHINE TRANSLATION HARD?

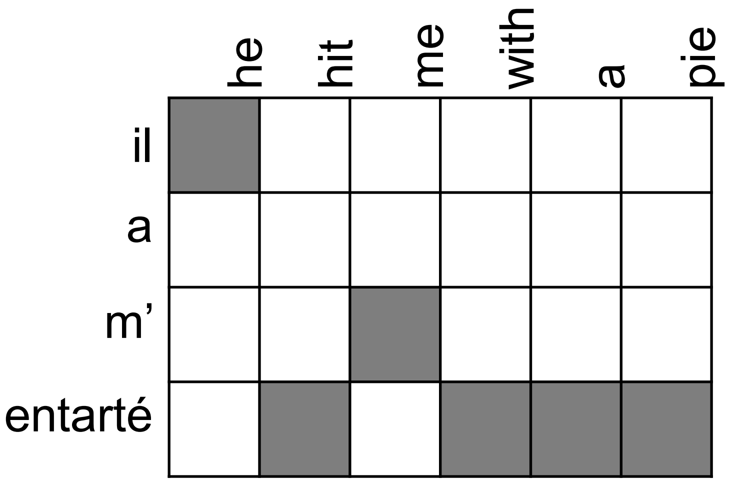


Examples from: “The Mathematics of Statistical Machine Translation: Parameter Estimation", Brown et al, 1993. http://www.aclweb.org/anthology/J93-2003

WHY IS MACHINE TRANSLATION HARD?

Some words are very fertile! Can map multiple words in the same sentence





SMT SYSTEMS ARE VERY COMPLEX

- Hundreds of important details
- Systems had many separately-designed subcomponents
- Lots of feature engineering
  - Need to design features to capture particular language phenomena
- Require compiling and maintaining extra resources
  - Like tables of equivalent phrases
- Lots of human effort to maintain
  - Repeated effort for each language pair!

MT EVALUATION

What do we need to evaluate?

- Correctness of the translation
- Fluency of the translation, appropriateness
- We need appropriate evaluation metrics

Automatic evaluation:

- Inexpensive, can be done on a large scale, but may not capture what we want to evaluate.

Human evaluation:

- Expensive, and not easily reproducible or comparable across evaluations (different judges, different questions, ...)

AUTOMATIC EVALUATION: BLUE

Evaluate candidate translations against several reference translations.

BLUE: Bilingual Evaluation Understudy Score

C1: It is a guide to action which ensures that the military always obeys the commands of the party.

C2: It is to insure the troops forever hearing the activity guidebook that party direct

R1: It is a guide to action that ensures that the military will forever heed Party commands.

R2: It is the guiding principle which guarantees the military forces always being under the command of the Party.

R3: It is the practical guide for the army always to heed the directions of the party.

The **BLUE** score is based on **n-gram** precisions:

- How many n-grams in the candidate translation occur also in one of the reference translation

AUTOMATIC EVALUATION: BLUE

Evaluate candidate translations against several reference translations.

BLUE: Bilingual Evaluation Understudy Score

C1: It is a guide to action which ensures that the military always obeys the commands of the party.

C2: It is to insure the troops forever hearing the activity guidebook that party direct

R1: It is a guide to action that ensures that the military will forever heed Party commands.

R2: It is the guiding principle which guarantees the military forces always being under the command of the Party.

R3: It is the practical guide for the army always to heed the directions of the party.

Unigram precision = 17/18

BLUE: ISSUE OF N-GRAM PRECISION

- What if some words are over-generated?
- An extreme example
  - Candidate: *the the the the the the the.*
  - Reference 1: *The cat is on the mat.*
  - Reference 2: *There is a cat on the mat.*
  - N-gram Precision: 7/7
- **Solution:**
  - reference word should be exhausted after it is matched.

BLUE: ISSUE OF N-GRAM PRECISION

- What if some words are just dropped?
- Another extreme example
  - Candidate: *the.*
  - Reference 1: *My mom likes the blue flowers.*
  - Reference 2: *My mother prefers the blue flowers.*
- N-gram Precision: 1/1
- **Solution:**
  - add a penalty if the candidate is too short.

BLEU

$$\text{BLEU} = \underbrace{(p_1 \cdot p_2 \cdot p_3 \cdot p_4)^{\frac{1}{4}}}_{} \underbrace{\max(1,\ e^{1-\frac{r}{c}})}_{\text{Brevity Penalty}}$$

Geometric Average

r = pick for each candidate in reference translation that is closest in length

Clipped N-gram precisions for N=1, 2, 3, 4

Brevity Penalty

c = length of the whole candidate translation corpus

- Ranges from 0.0 to 1.0, but usually shown multiplied by 100
- An increase of +1.0 BLEU is usually a conference paper
- MT systems usually score in the 10s to 30s
- Human translators usually score in the 70s and 80s

BLUE ADVANTAGES

- Quick and inexpensive to calculate
- It is easy to understand
- It is language independent
- It correlates highly with human evaluation

HUMAN EVALUATION

We want to know whether the translation is **“good” and accurate** of the original.

- Ask humans to judge the **fluency** and the **adequacy** of the translation
  - (e.g. on a scale of 1 to 5)

- Correlated with fluency is accuracy on **cloze task**:
  - Give raters the sentence with one word replaced by blank.
  - Ask raters to guess the missing word in the blank.

- Similar to adequacy is **informativeness**
  - Can you use the translation to perform some task
  - (e.g. answer multiple-choice questions about the text)

REFERENCES