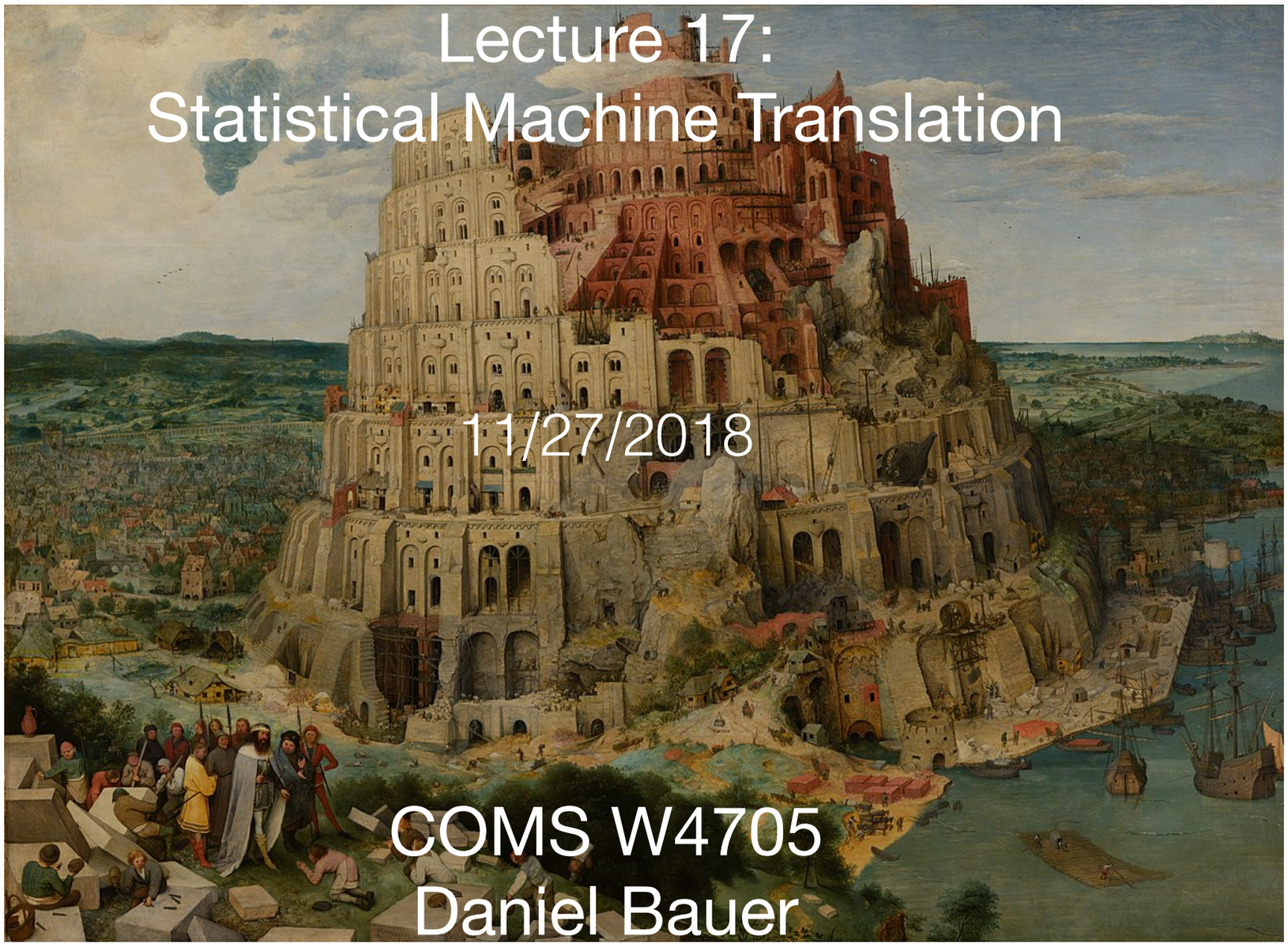
Natural Language Processing

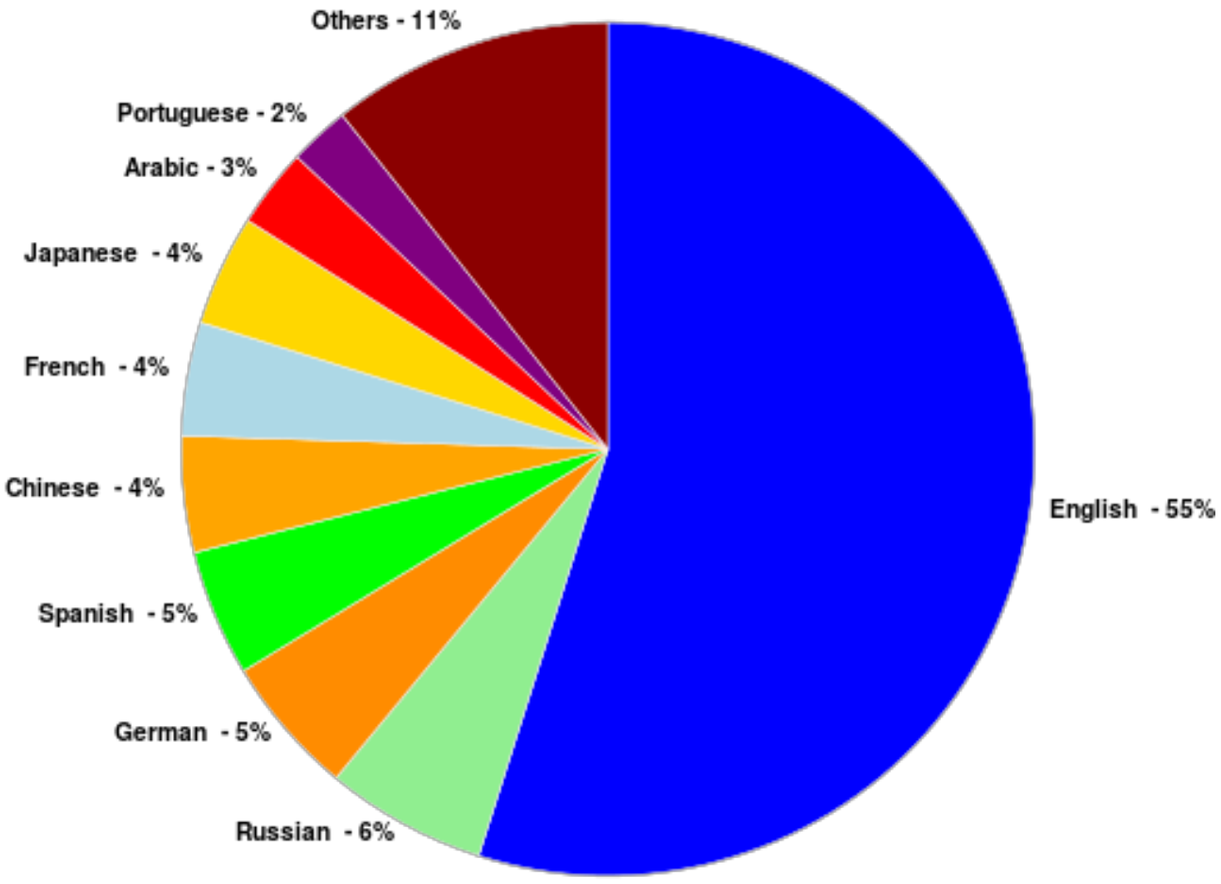


Lecture 17:
Statistical Machine Translation

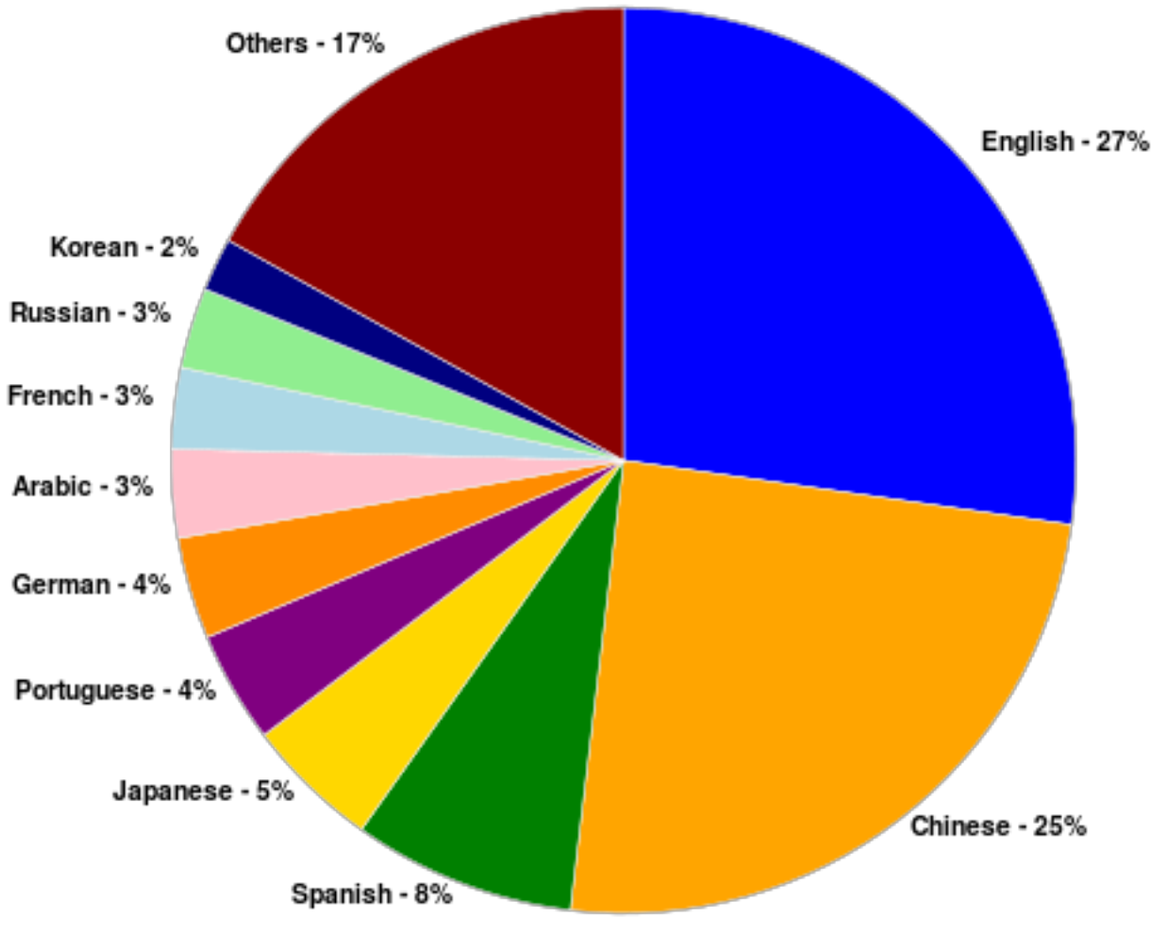
11/27/2018

COMS W4705
Daniel Bauer

Languages on the Internet



Content Languages for Websites



Internet Users by Language

Some Challenges for MT

- Systematic differences in languages
  - Morphology, syntax, ...
- Ambiguity in translation.
  - Lexical ambiguity, syntactic ambiguity, ...

Language Differences
Syntax

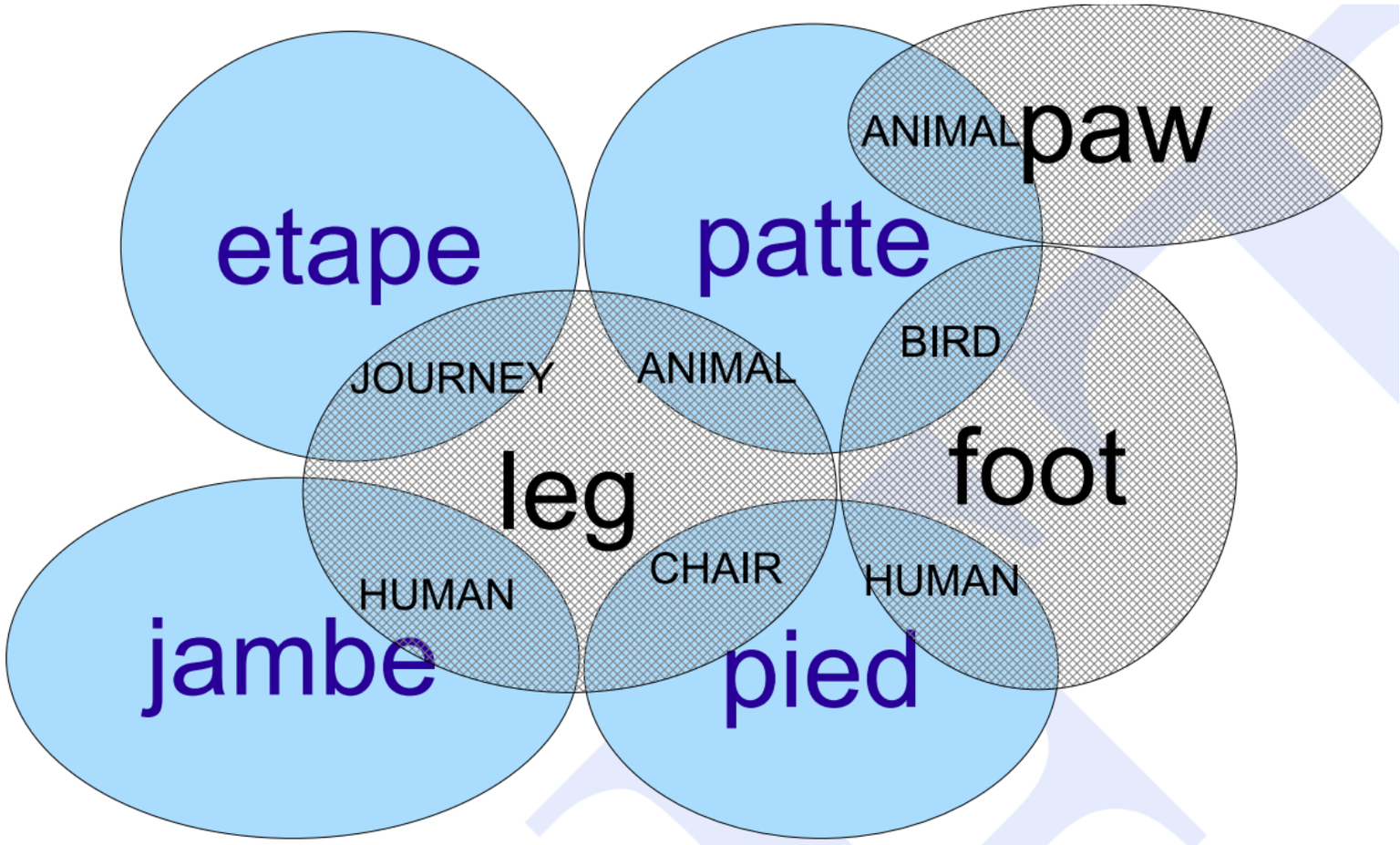
- Sentence Word order:
  - (S)ubject (V)erb (O)bject: English, Mandarin
  - VSO: Irish, Classical Arabic
  - SOV: Hindi, Japanese

JP: *Torako ga nezumi o mimashita*
Torako-subj mouse-obj saw
EN: *Torako saw a mouse*

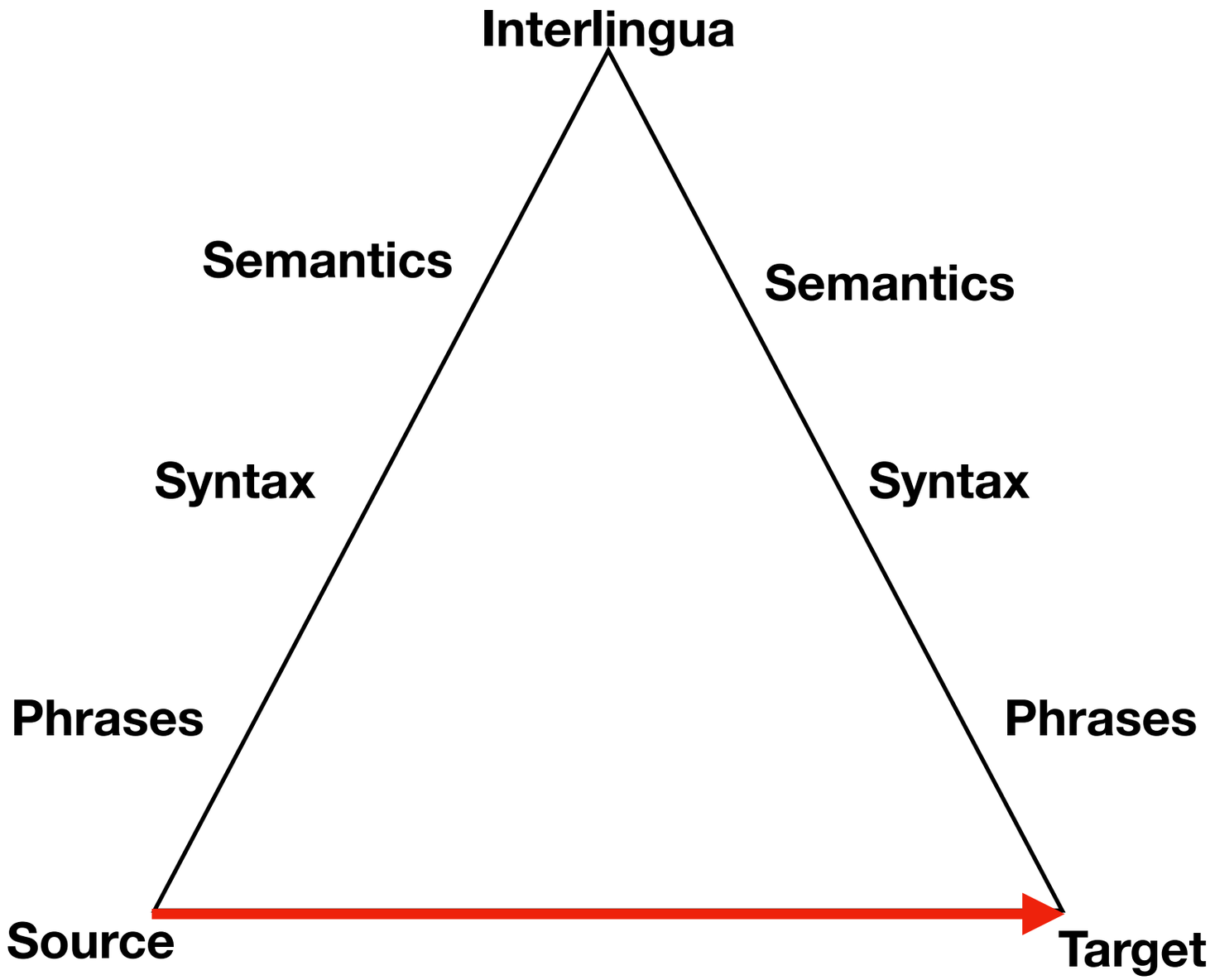
- Word order in phrases (adjective modifiers):
  - EN: *the blue house,* FR: *la maison bleue*
- Prepositions and case marking:

JP: *Torako-ni*
EN: *to Torako*

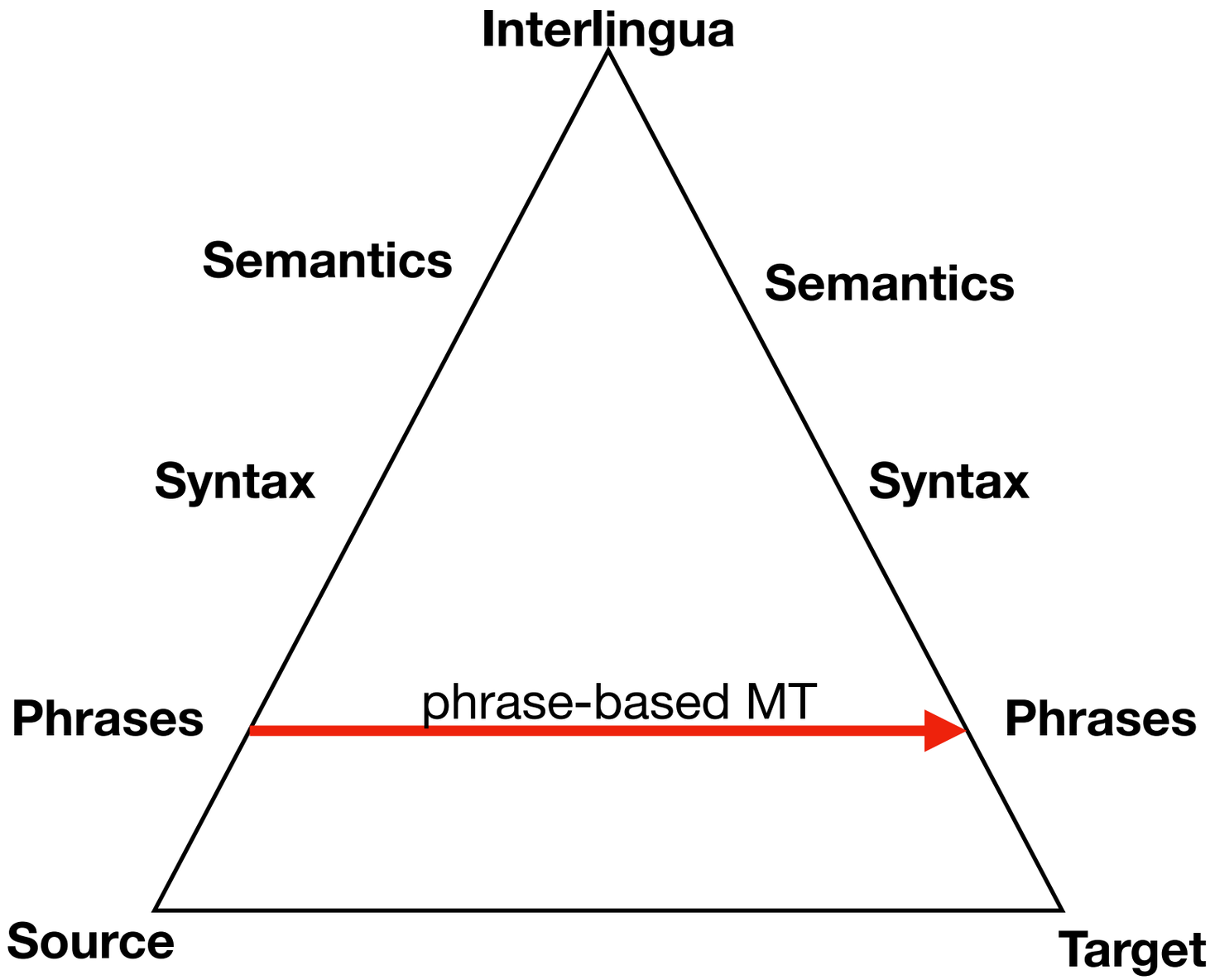
Lexical Divergence



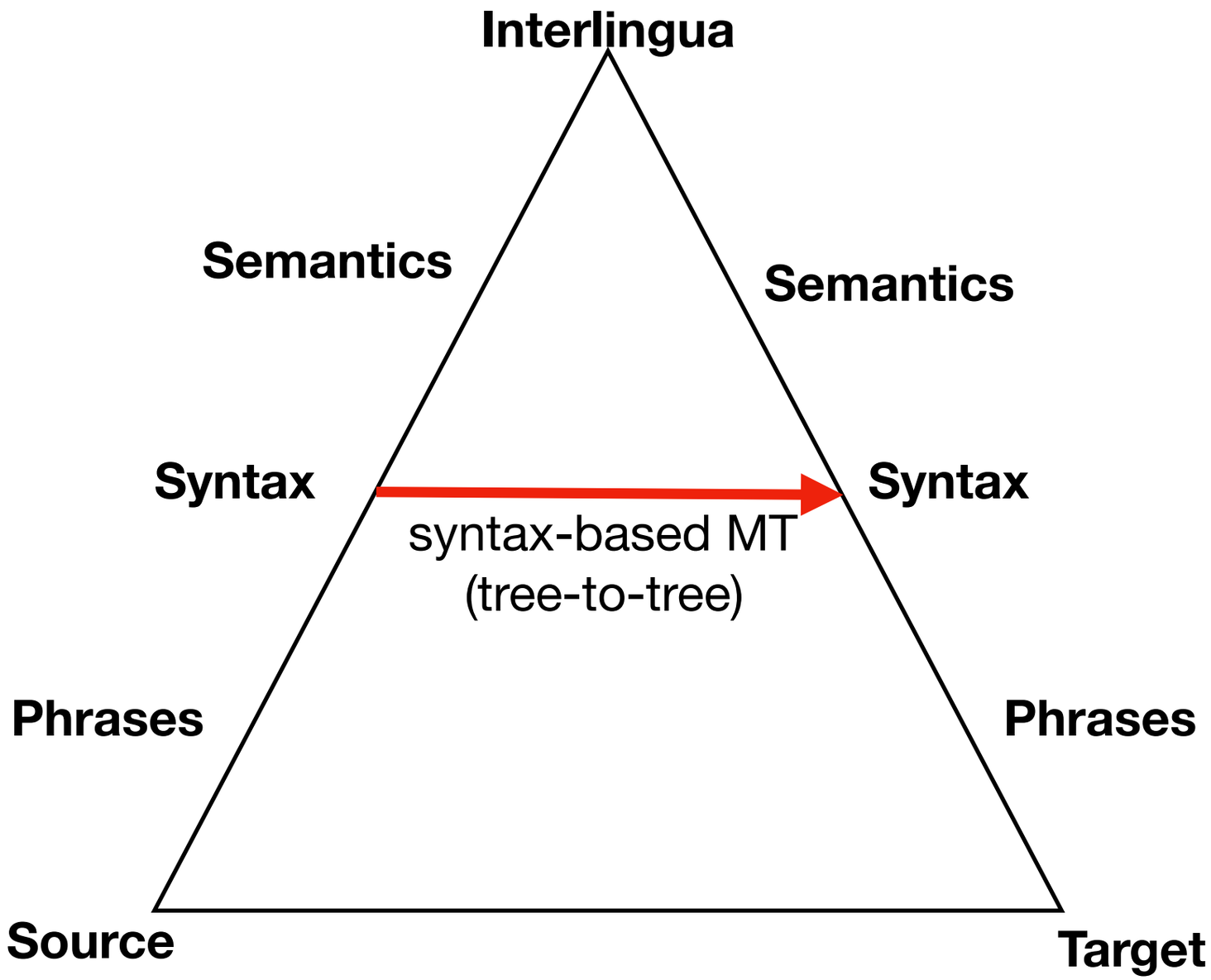
Vauquois Triangle



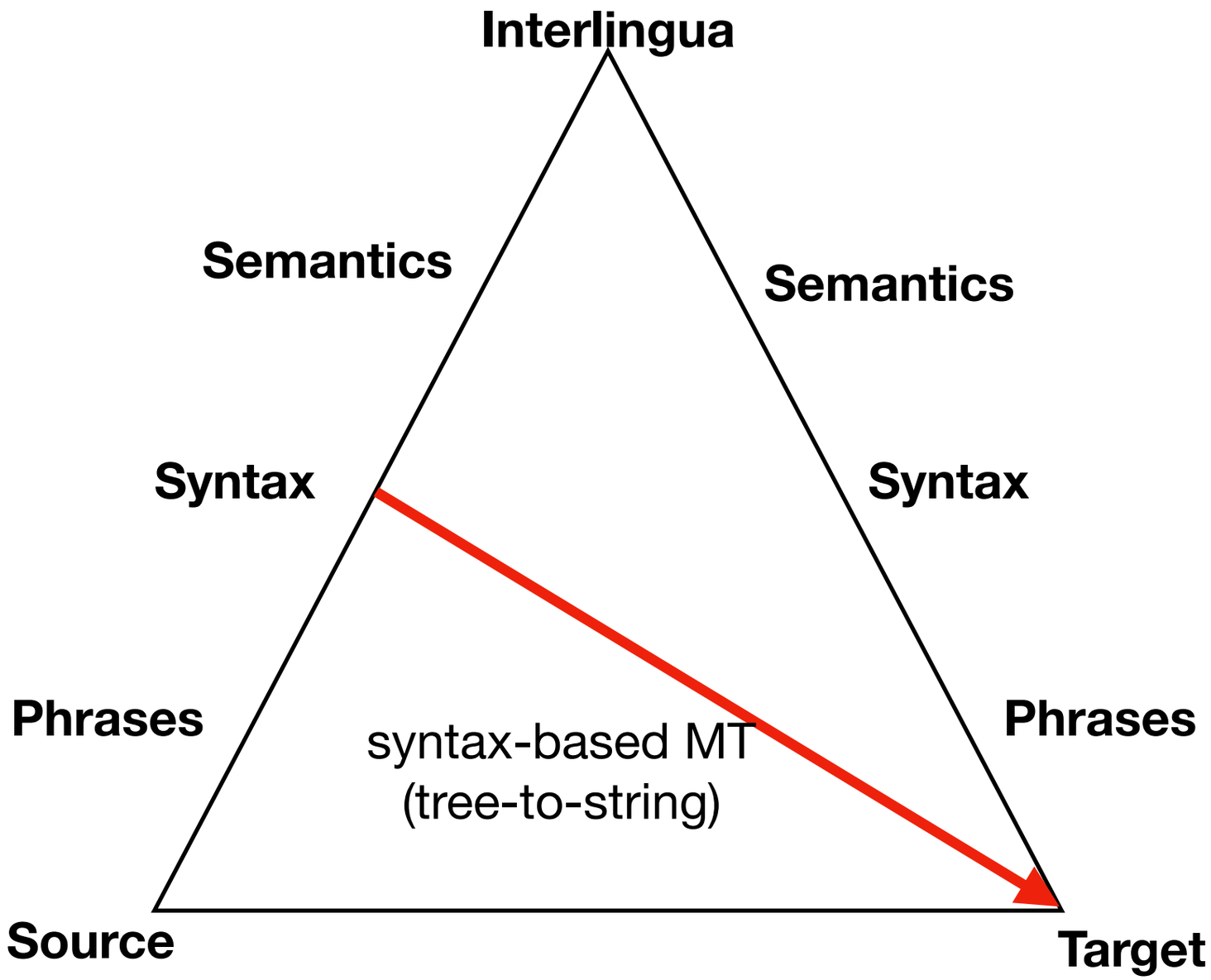
Vauquois Triangle



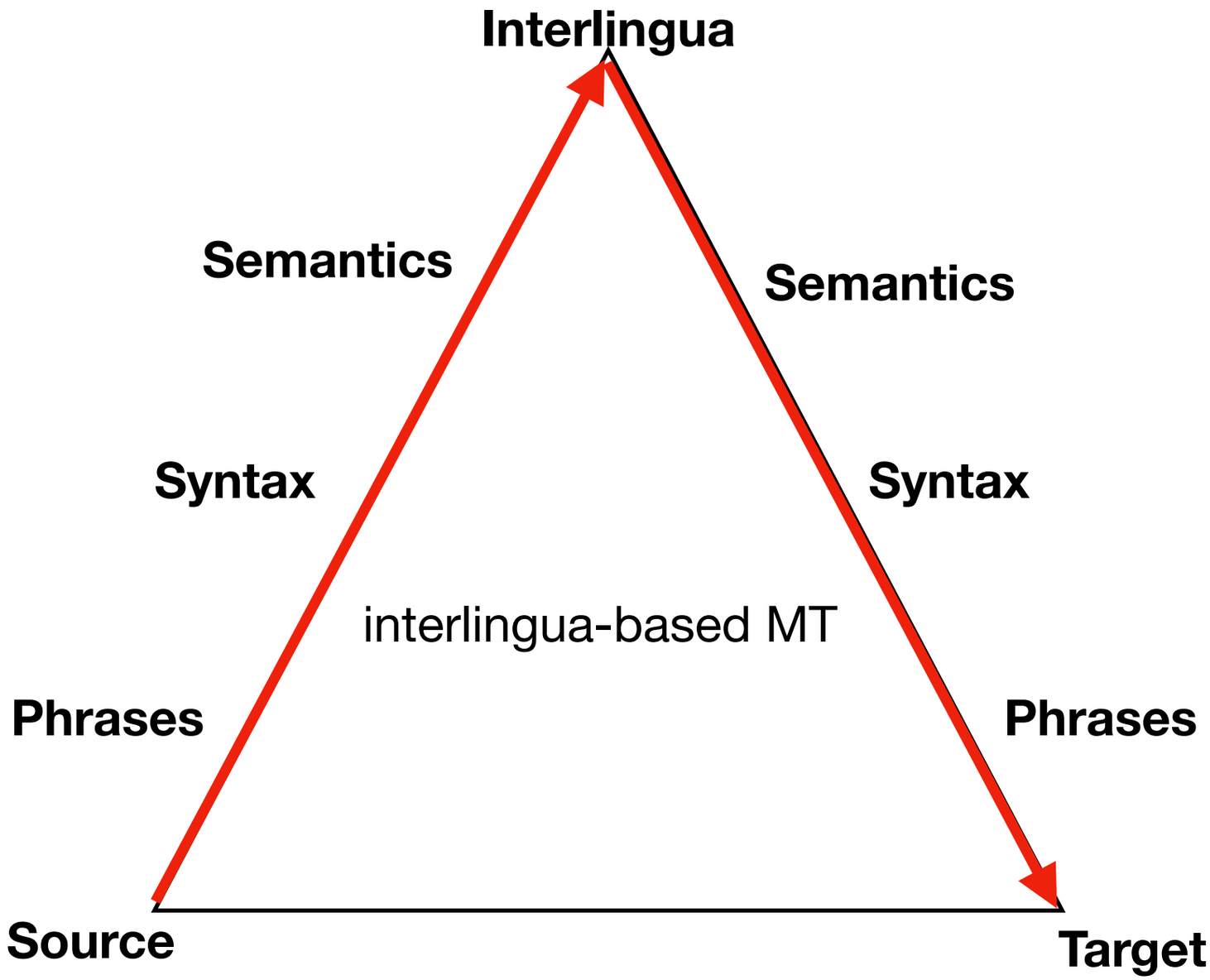
Vauquois Triangle



Vauquois Triangle



Vauquois Triangle



Overview

- Statistical MT:
  - Alignments, IBM models (noisy channel model)
  - Phrase-based MT
  - (Syntax-based MT)
- MT Evaluation
- Neural MT (encoder-decoder model, seq2seq), on Wednesday

Faithfulness and Fluency

- Good translation needs to be:
    - Faithful: Target sentence should have the same *content* as the source text.
    - Fluent: Target text should be grammatical / natural / fluent in the target language.
- There is often a trade-off between these two factors.

Fluency

- **Fluency:** Assume a translation system produces the following output candidates:

    1. *That car near crash to me.*
    2. *That car almost hit me.*

    - Which sentence would you choose? How would we model this?

- Can use any mono-lingual language model!

Faithfulness

- **Faitfulness:**
  - Source sentence:

    Maria no dió una bofetada a la bruja verde

  - Which sentence is the most faithful output:

    1. *Mary didn’t slap the green witch*
    2. *Mary not give a slap to the witch green*
    3. *The green witch didn’t slap Mary*
    4. *Mary slapped the green witch*

  - How would we model faithfulness?

MT as Decipherment

"One naturally wonders if the problem of translation could conceivably be treated as a problem in cryptography.
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, "Translation", 1955)

Slide inspired by Dragomir Radev & Kevin Knight

Noisy Channel Model for MT

EN: "*Mary did not slap the green witch*"

P(E)

P(F | E)

direction
of translation

ES: "*Maria no dió una bofetada a la bruja verde*"

- Goal: Given an observation in the **source language** (F), figure out what was said in the **target language** (E).
- Fluency is modeled by $\mathrm{P(E)}$.
  Faithfulness is modeled by $\mathrm{P(F|E)}$.
- Apply Baye's rule: $\mathrm{P(E|F)} \propto \mathrm{P(E)\ P(F|E)}$

Faithfulness and Fluency Example

FR: une fleur rouge

	P(e)	P(f|e)	P(e) * P(f|e)
1. a flower red	low	high	low
1. red flower a	low	high	low
3. flower red a	low	high	low
4. a red dog	high	low	low
5. dog cat mouse	low	low	low
6. a red flower	high	high	high

Parallel Corpora

The Rosetta Stone

- Carved in 196 BC in Egypt
- Identical text written in three scripts:
  - Hieroglyphics, Demotic, and Greek
- Hieroglyphics decipherd by J-F. Champollion in 1822.

Two Egyptian scripts: Hieroglyphics and Demotic

http://www.ancientegypt.co.uk/writing/rosetta.html
Slide inspired by Dragomir Radev & Kevin Knight

Modern Parallel Corpora

- **Hansards**: Proceedings of the Canadian Parliament. French/English, ~1m sentence pairs.

- **EUROPARL**: Proceedings of the European Parliament. Translated into 21 EU languages:
    - Romanic (French, Italian, Spanish, Portuguese, Romanian), Germanic (English, Dutch, German, Danish, Swedish), Slavik (Bulgarian, Czech, Polish, Slovak, Slovene), Finni-Ugric (Finnish, Hungarian, Estonian), Baltic (Latvian, Lithuanian), and Greek

- between 400k and 1.5m sentence pairs per language pair. **(http://www.statmt.org/europarl/)**

Europarl Example

Danish: det er næsten en personlig rekord for mig dette efterår .
German: das ist für mich fast persönlicher rekord in diesem herbst .
Greek: πρόκειται για το προσωπικό μου ρεκόρ αυτό το φθινόπωρο .
English that is almost a personal record for me this autumn !
Spanish: es la mejor marca que he alcanzado este otoño .
Finnish: se on melkein minun ennätykseni tänä syksynä !
French: c ' est pratiquement un record personnel pour moi , cet automne !
Italian: e ' quasi il mio record personale dell ' autunno .
Dutch: dit is haast een persoonlijk record deze herfst .
Portuguese: é quase o meu recorde pessoal deste semestre !
Swedish: det är nästan personligt rekord för mig denna höst !

(Philipp Koehn, "Europarl: A Parallel Corpus for Statistical Machine Translation", http://homepages.inf.ed.ac.uk/pkoehn/publications/europarl-mtsummit05.pdf)

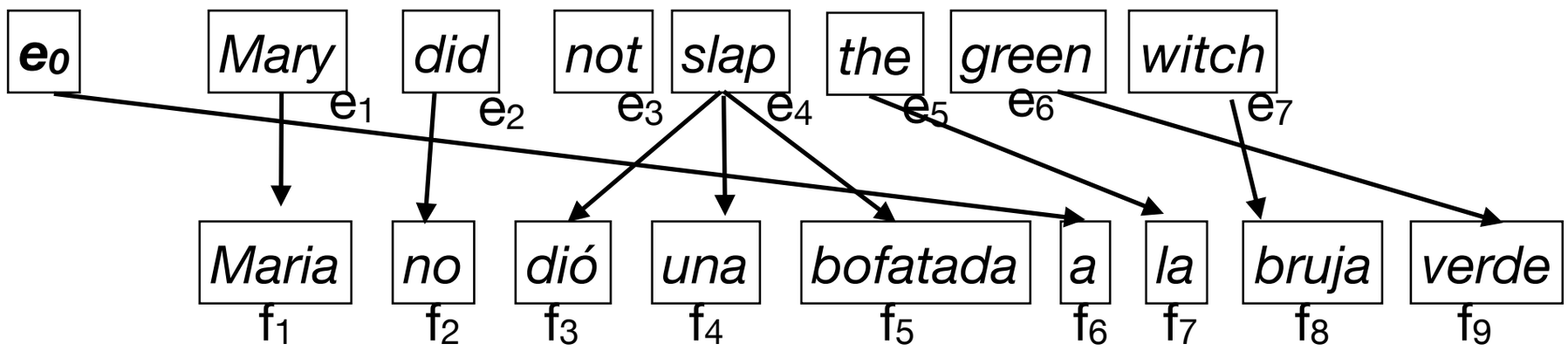
Word Alignments

- First step: Figure out word-to-word translations by aligning words between parallel sentences (bitext) in the corpus.

	Maria	*no*	*dió*	*uno*	*bofatada*	*a*	*la*	*bruja*	*verde*
Mary									
did									
not									
slap									
the									
green									
witch									

Word Alignments: IBM Model 2

- Simplifying assumption:
  alignments are one-to-many, i.e.
  each *f* word originates from exactly one *e* word (not many-to-one, many-to-many, ...)



- Use alignment variables $a_1 \ldots a_m$ to indicate the position that each word in *f* is aligned to.

$$a_{1, \ldots, a_9} = \langle 1, 2, 4, 4, 4, 0, 5, 7, 6 \rangle$$

IBM Model 2

- We can use the alignment variables to defined a "mini translation model"

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

- Decoding problem: find

$$\arg\max_{E} P(E)P(F|E)$$

- This is not used for translation in practice, but forms the basis for computing alignments.

- We want to model the conditional probability

$$P(F, A|E) = P(f_1 \dots f_m, a_1 \dots a_m | e_1 \dots e_l, m)$$

IBM Model 2

- Two parameters
  - $t(f|e)$ is the probability of generating word f from e.
  - $q(j|i,l,m)$ is the probability that alignment variable i takes value j (given sentence length m and l).

- Define the conditional probability for the target sentence and alignments as

$$P(f_1 \dots f_m, a_1 \dots a_m | e_1 \dots e_l, m) = \prod_{i=1}^{m} q(a_i|i,l,m) t(f_i|e_{a_i})$$

- What are the independence assumptions in this model?

Computing Alignments

- The alignments computed by IBM models form the basis of phrase-based translation.

- We want to compute the optimal alignments given a sentence pair:

$$\arg \max_{a_1 \ldots a_m} P(a_1 \ldots a_m | f_1 \ldots f_m, e_1 \ldots e_l, m)$$

- Because the a_i are all independent:

$$a_i = \arg \max_{j \in \{0 \ldots l\}} q(j|i,l,m) t(f_i, e_j) \qquad \text{for i = 1...m}$$

Training the Model
Fully Observable Case

- In reality, we only get (E/F) pairs for training. We do not know the alignments ("partially observed data").

- Assume we *did* know the alignments. Then we could estimate

$$t(f|e) = \frac{count(e, f)}{count(e)}$$

$$q(j|i, l, m) = \frac{count(j|i, l, m)}{count(i, l, m)}$$

- How do we deal with the partially observed case?

Observed Counts

- Initialize expected counts to 0.
- For each sentence pair k=1...n
  - For i=1...m_k:
    - For j = 0...l_k:

$$\delta(k,i,j) = \begin{cases} 1 \text{ if } a_i = j \\ 0 \text{ otherwise} \end{cases}$$

$$count(e_j, f_i) \leftarrow count(e_j, f_i) + \delta(k,i,j)$$

$$count(e_j) \leftarrow count(e_j) + \delta(k,i,j)$$

$$count(j|i,l,m) \leftarrow count(j|i,l,m) + \delta(k,i,j)$$

$$count(i,l,m) \leftarrow count(i,l,m) + \delta(k,i,j)$$

Training the Model
Partially Observable Case

- Expectation Maximization (EM) Algorithm.

- Start with uniform distribution for the t and q tables, then iteratively refine them.

- At each iteration

  - first compute **expected counts** based on the current t and q values and the data

    Expectation Step

    - $\text{count}(e)$, $\text{count}(e,f)$, $\text{count}(j \mid i,l,m)$, $\text{count}(i,l,m)$

- Then use the expected counts to re-estimate t and q.

Maximization Step

Expected Counts

- Initialize expected counts to 0.
- For each sentence pair k=1...n
  - For i=1...m_k:
    - For j = 0...l_k:

$$\delta(k,i,j) = \frac{q(j|i,l_k,m_k))t(f_i|e_j)}{\sum_{j=0}^{l} q(j|i,l_k,m_k)t(f_i|e_j)}$$

$$count(e_j, f_i) \leftarrow count(e_j, f_i) + \delta(k,i,j)$$

$$count(e_j) \leftarrow count(e_j) + \delta(k,i,j)$$

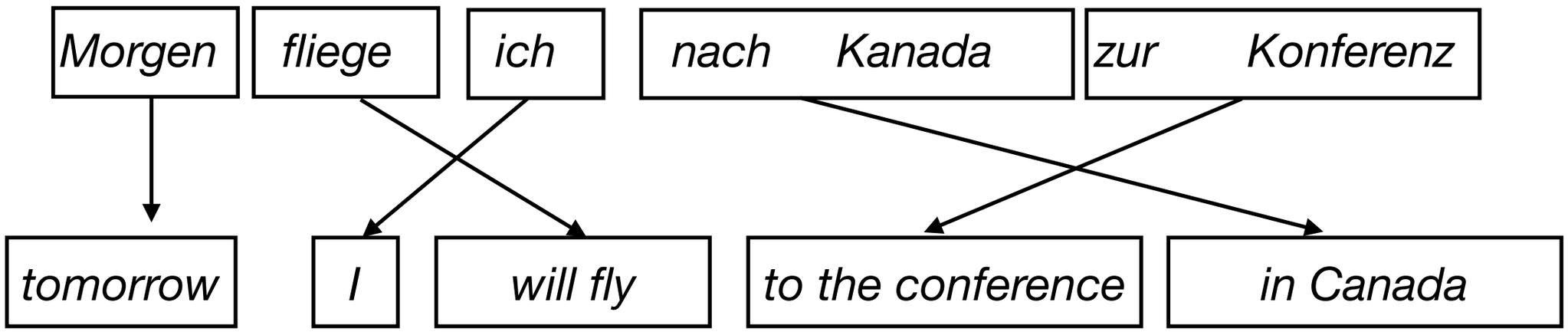
$$count(j|i,l,m) \leftarrow count(j|i,l,m) + \delta(k,i,j)$$

$$count(i,l,m) \leftarrow count(i,l,m) + \delta(k,i,j)$$

Phrase-Based Statistical MT

- Popular, state-of-the-art approach prior to neural MT.

- Input segmented into phrases (a phrase is any sequence of words).



- Each phrase-pair has a translation probability ("phrase table")
  P(to the conference | zur Konferenz)
  P(for the conference | zur Konference)

- Probabilistically create output phrases, then reorder them (distortion parameters).

(Koehn, Och & Marcu "Statistical Phrase-Based Translation", 2003)

Creating Phrase Alignments

- Word alignments (IBM Models) are one-to-many.
- Phrase alignments are (many-to-many).
- Get alignments in each direction (E→F, F→E), then merge the two alignments into one minimal alignment.
- Extract **all** phrases compatible with the minimal alignment.

Word Alignment Induced Phrases

	Maria	*no*	*dió*	*uno*	*bofatada*	*a*	*la*	*bruja*	*verde*
Mary	■								
did						■			
not		■							
slap			■	■	■				
the							■		
green									■
witch								■	

English → Spanish

Word Alignment Induced Phrases

	Maria	*no*	*dió*	*uno*	*bofatada*	*a*	*la*	*bruja*	*verde*
Mary	■								
did		■							
not		■							
slap					■				
the							■		
green									■
witch								■	

Spanish → English

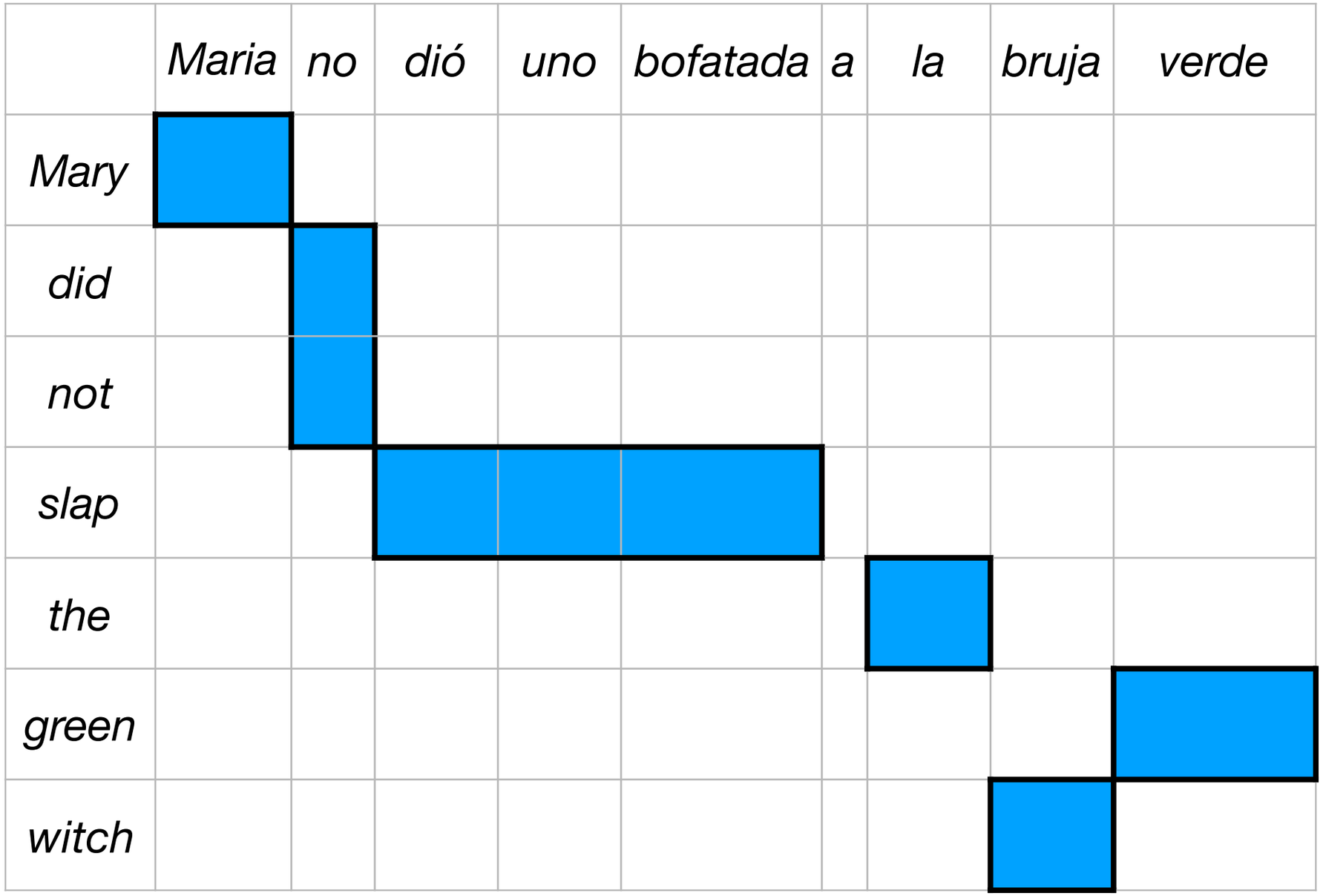
Word Alignment Induced Phrases

	Maria	*no*	*dió*	*uno*	*bofatada*	*a*	*la*	*bruja*	*verde*
Mary	■								
did									
not		■							
slap					■				
the							■		
green									■
witch								■	

Merged into Spanish ↔ English minimal alignment.

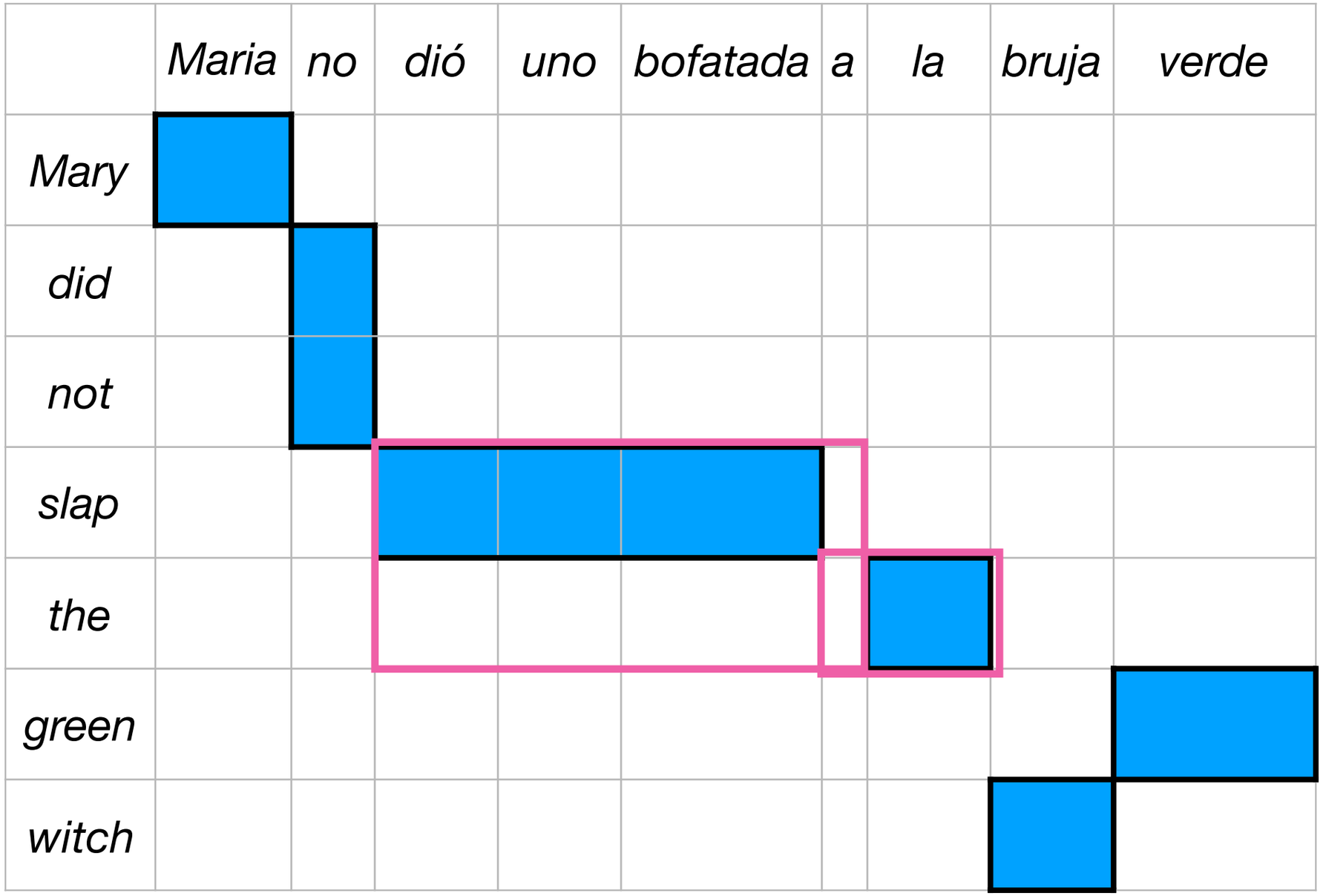
(intersection)

Word Alignment Induced Phrases



(Maria, Mary) (no, did not) (slap, dió una bofetada) (la, the) (bruja, witch) (verde, green)

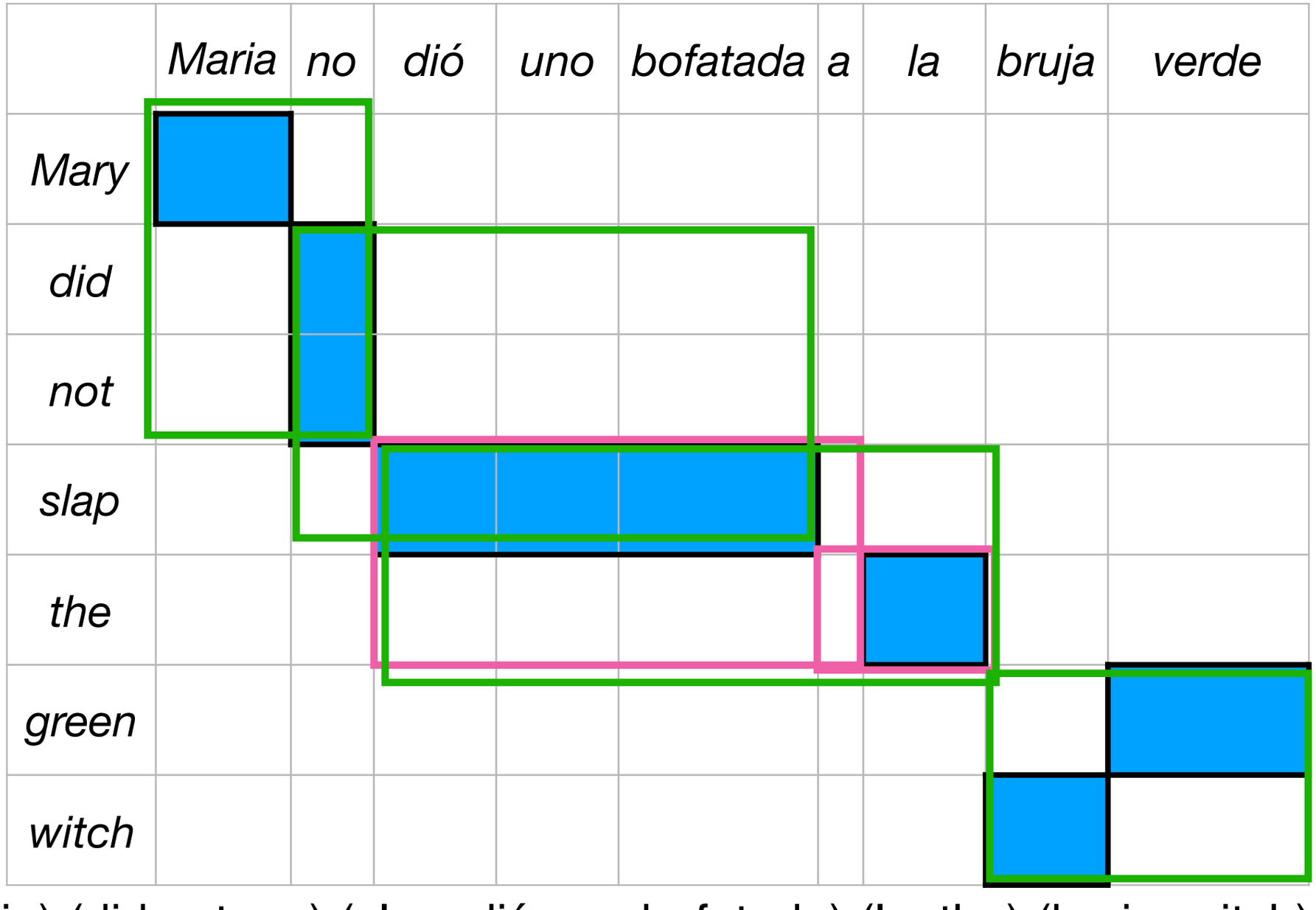
Word Alignment Induced Phrases



(Maria, Mary) (no, did not) (slap, dió una bofetada) (la, the) (bruja, witch) (verde, green)

(the, a la)(slap the, dió una bofetada a)

Word Alignment Induced Phrases

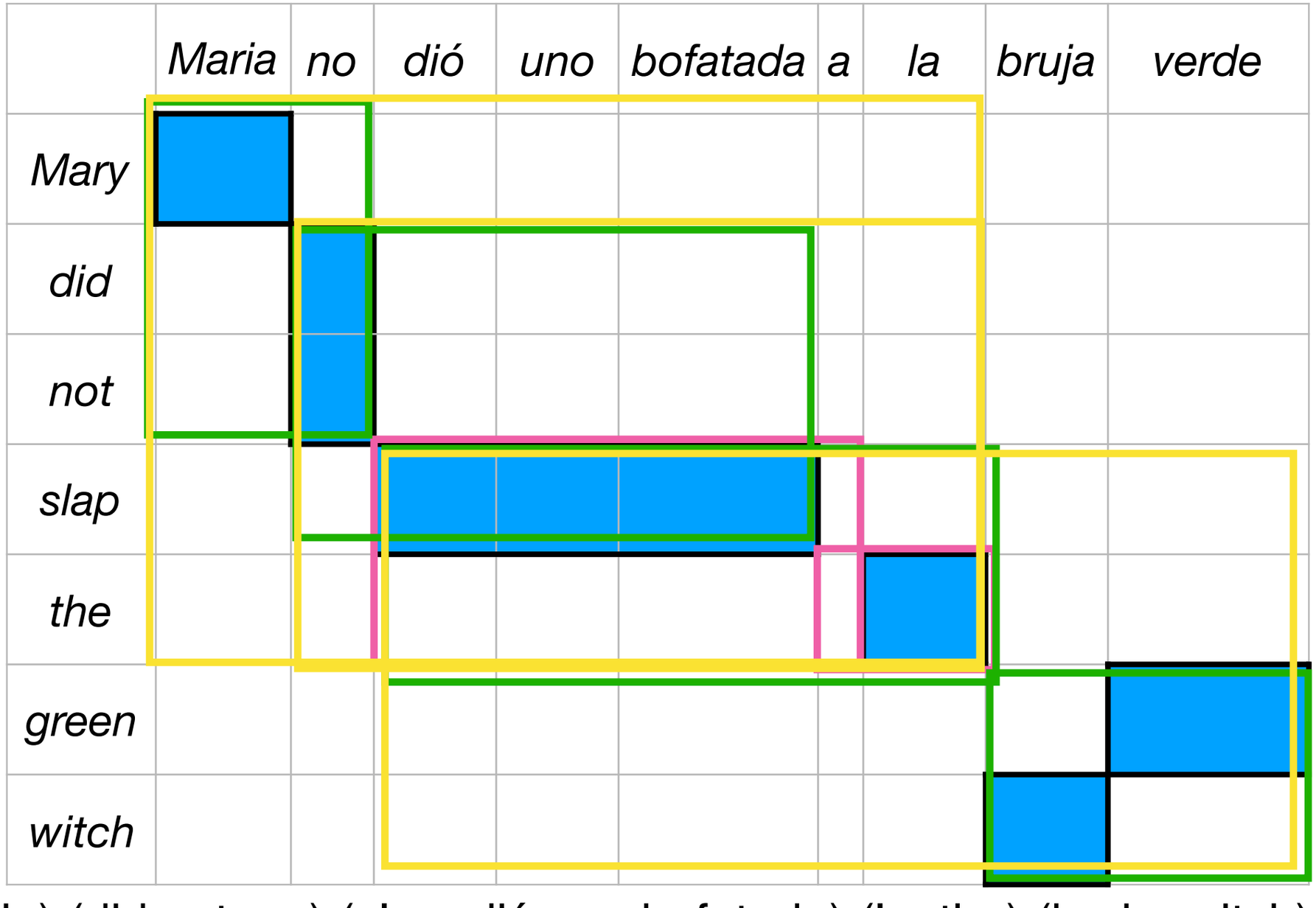


(Mary, Maria) (did not, no) (slap, dió una bofetada) (la, the) (bruja, witch) (verde, green)

(the, a la)(slap the, dió una bofetada a) (Mary did not, Maria no) ...

(did not slap, dió una bofetada)(slap the, dió una bofetada a la)(green witch, bruja verde)

Word Alignment Induced Phrases



(Mary, Maria) (did not, no) (slap, dió una bofetada) (la, the) (bruja, witch) (verde, green)

(the, a la)(slap the, dió una bofetada a) (Mary did not, Maria no) ...

(did not slap, dió una bofetada)(slap the, dió una bofetada a la)(green witch, bruja verde)

Advantages of Phrase-Based MT

- Many-to-many mappings can handle non-compositional phrases.

- Local context is useful for disambiguating
  EN: "*Interest rate*"->DE: "*Zinssatz*"
  EN: "*Interest in*" -> DE:"*Interesse an*"

- With more data it becomes more likely that phrases have been seen before (even entire sentences).

  - Phrase tables are also common in computer-assisted human translation.

MT Evaluation

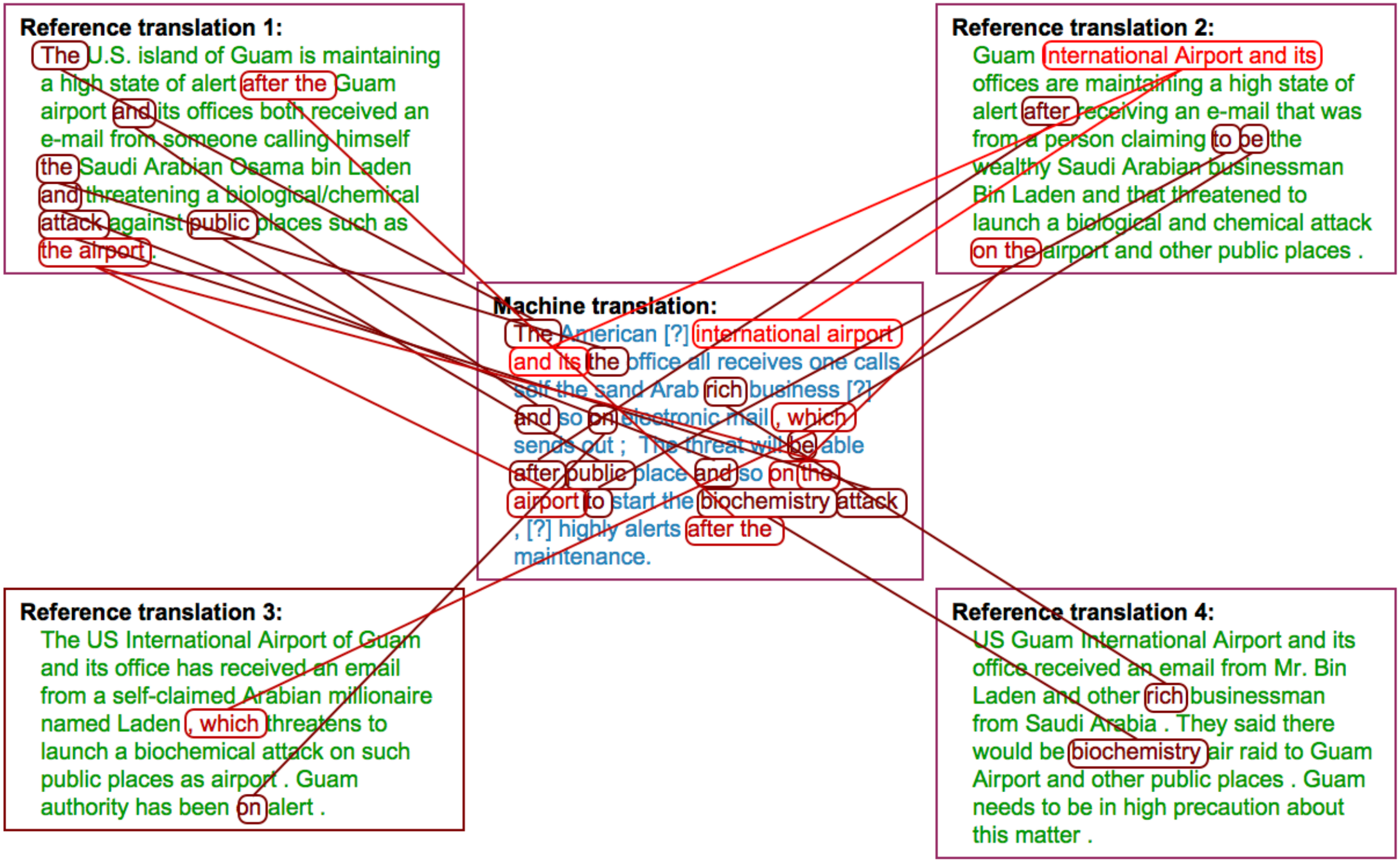
- MT Evaluation is notoriously difficult!
- No single correct output (typically use multiple reference translations).
- Need to evaluate faithfulness and fluency, but both are subjective.
- Dumb machines vs. slow humans.
- Wide range of different automatic metrics (BLEU, NIST, TER, METEOR)

BLEU Metric

- **B**i**L**ingual **E**valuation **U**nderstudy
- Modified n-gram precision with length penalty. Recall is ignored.
- Quick, inexpensive, and language independent.
- Correlates highly with human evaluations.
- But: Bias against synonyms and inflectional variations. Penalizes variations in word-order between languages in different families.

(Papineni, et al. 2002. "BLEU: A method for automatic evaluation of machine translation.")

Multiple Reference Translations



Slide from Bonnie Dorr

Computing BLEU

- Test sentence:
  Colorless green ideas sleep furiously

- Reference translations:
  all dull jade ideas sleep irately
  drab emerald concepts sleep furiously
  colorless immature thoughts nap angrily

Computing BLEU

- Test sentence:
  Colorless *green* ***ideas sleep furiously***

- Reference translations:
  all dull jade ***ideas*** *sleep irately*
  drab emerald concepts ***sleep furiously***
  colorless *immature thoughts nap angrily*

Unigram precision: 4/5

Computing BLEU

- Test sentence:
  *Colorless green **ideas sleep** furiously*
  sleep furiously
- Reference translations:
  *all dull jade **ideas sleep** irately*
  *drab emerald concepts **sleep furiously***
  colorless immature thoughts nap angrily

Unigram precision: 4/5

Bigram precision: 2/4

Computing BLEU

- Test sentence:
  *Colorless green **ideas sleep** furiously*
  sleep furiously
- Reference translations:
  *all dull jade **ideas sleep** irately*
  *drab emerald concepts **sleep furiously***
  colorless immature thoughts nap angrily

Unigram precision: 4/5 =0.8

Bigram precision: 2/4 =0.5

BLEU score = $(a_1 \times a_2 \times \ldots \times a_n)^{1/n}$

$= (0.8 \times 0.5)^{1/2} = 0.6325$ → **63.25**

BLEU: Brevity Penalty

- BLEU is precision based. Dropped words are not penalized.

- Instead, a **brevity penalty** is used for translations that are shorter than the reference translations.

- Let c be the length of the candidate translation and r be the length of the reference translation that has the closest length.

$$BP = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases}$$

$$BLEU = BP \cdot exp\left(\sum_{n=1}^{N} \frac{1}{n} \log \text{ precision}_n\right)$$

Acknowledgments

- Some slides adapted from Kathy McKeown, Dragomir Radev, Dan Jurafsky, Kevin Knight, Bonnie Dorr.

- Presentation of IBM M2 adapted from Michael Collins