Introduction to Natural Language Processing

Machine Translation

- Jurafsky, D. and Martin, J. H. (2021): Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition. Third Edition. Chapter 10:
- Manning, C. D. and Schütze, H. (1999): Foundations of Statistical Natural Language Processing. MIT Press: Cambridge, Massachusetts. Chapters 2.1, 2.2, 6.
- Koehn, P. (2009). Statistical machine translation. Cambridge University Press.

PLAN OF THE LECTURE

- Machine Translation (MT) Task
- Rule-based MT Models
- Statistical MT Models
- Neural MT Models

The task of machine translation

- Machine Translation (MT) is the task of translating a sentence x from one language (the source language) to a 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`

http://web.stanford.edu/class/cs224n/slides/cs224n-2020-lecture08-nmt.pdf

Early machine translation

- Machine Translation research began in the early 1950s.
- Russian → English (motivated by the Cold War!)
- Systems were mostly rule-based, using a bilingual dictionary to map Russian words to their English counterpart

http://web.stanford.edu/class/cs224n/slides/cs224n-2020-lecture08-nmt.pdf

Early machine translation

- Machine Translation research began in the early 1950s.
- Russian → English (motivated by the Cold War!)
- Systems were mostly rule-based, using a bilingual dictionary to map Russian words to their English counterpart

Early results from translating English into Russian and back to English:

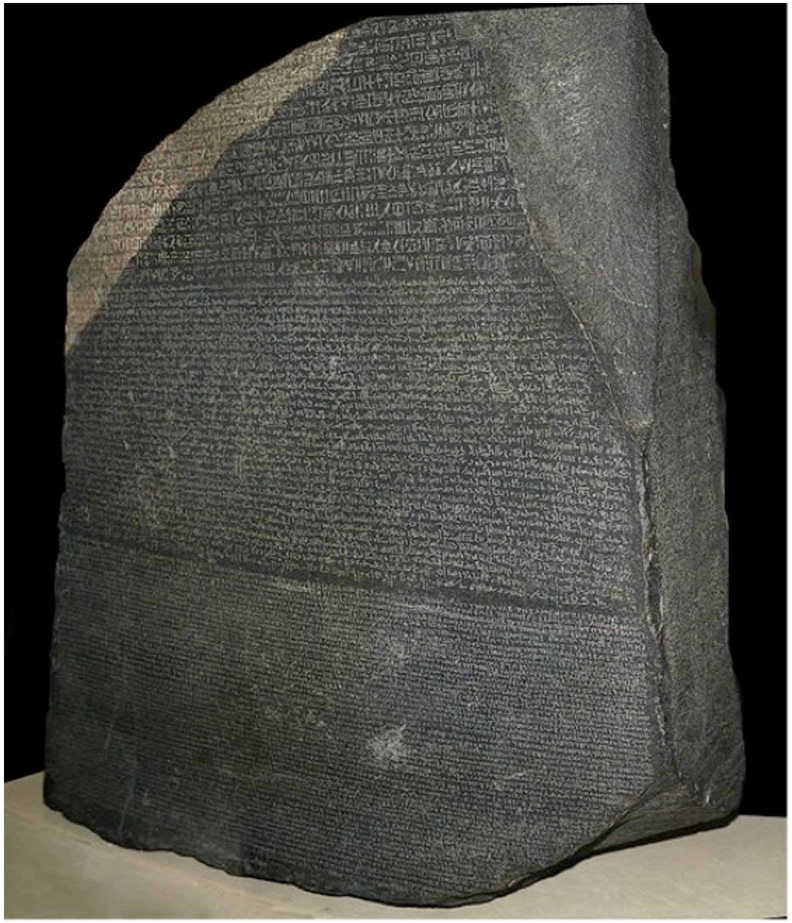
- The spirit is willing but the flesh is weak
- The vodka is good but the meat is rotten

http://web.stanford.edu/class/cs224n/slides/cs224n-2020-lecture08-nmt.pdf

Parallel Corpus: Training resource for MT

Most popular:

- EuroParl: European parliament protocols in 11 languages
- Hansards: Canadian Parliament protocols in French and English
- Software manuals (KDE, Open Office …)
- Parallel webpages



Rosetta stone (196 BC): Greek-Egyptian-Demotic

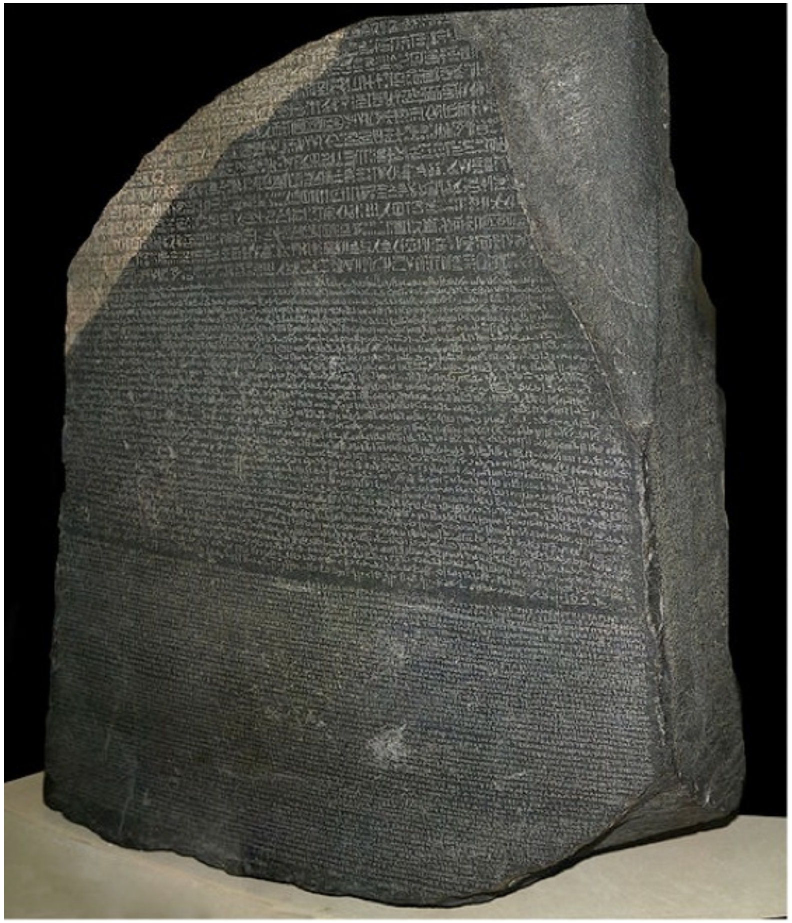
Parallel Corpus: Training resource for MT

Most popular:

- EuroParl: European parliament protocols in 11 languages
- Hansards: Canadian Parliament protocols in French and English
- Software manuals (KDE, Open Office …)
- Parallel webpages

Usually we assume that we have a **sentence-aligned** parallel corpus.

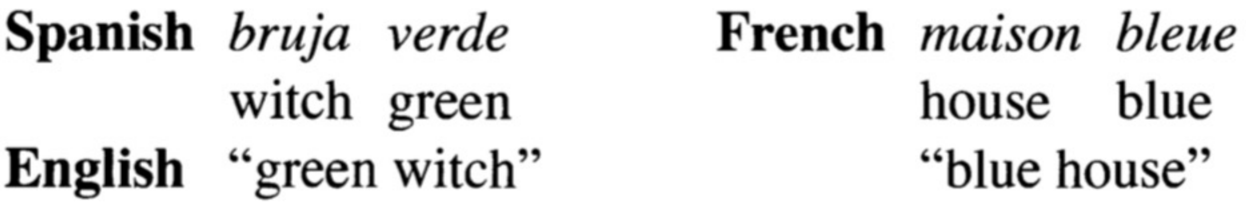
- there are methods to get to aligned sentences from aligned documents
- there are methods to extract parallel sentences from comparable corpora



Rosetta stone (196 BC): Greek-Egyptian-Demotic

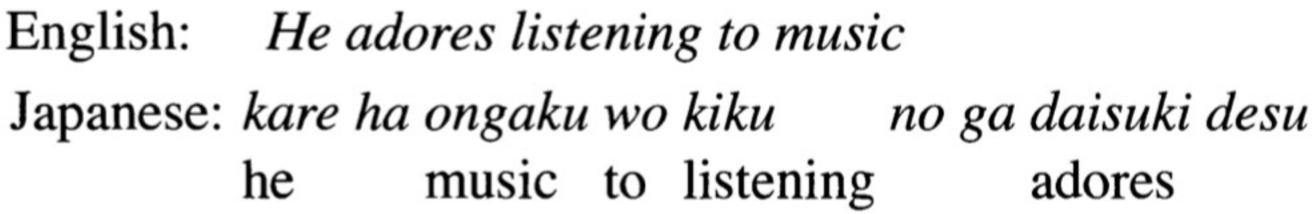
Why machine translation is hard?

- Languages are structurally very different:
  - Word order



Spanish	*bruja*	*verde*	**French**	*maison*	*bleue*
	witch	green		house	blue
English	"green witch"			"blue house"	

  - Syntax (e.g. SVO vs SOV vs VSO languages)

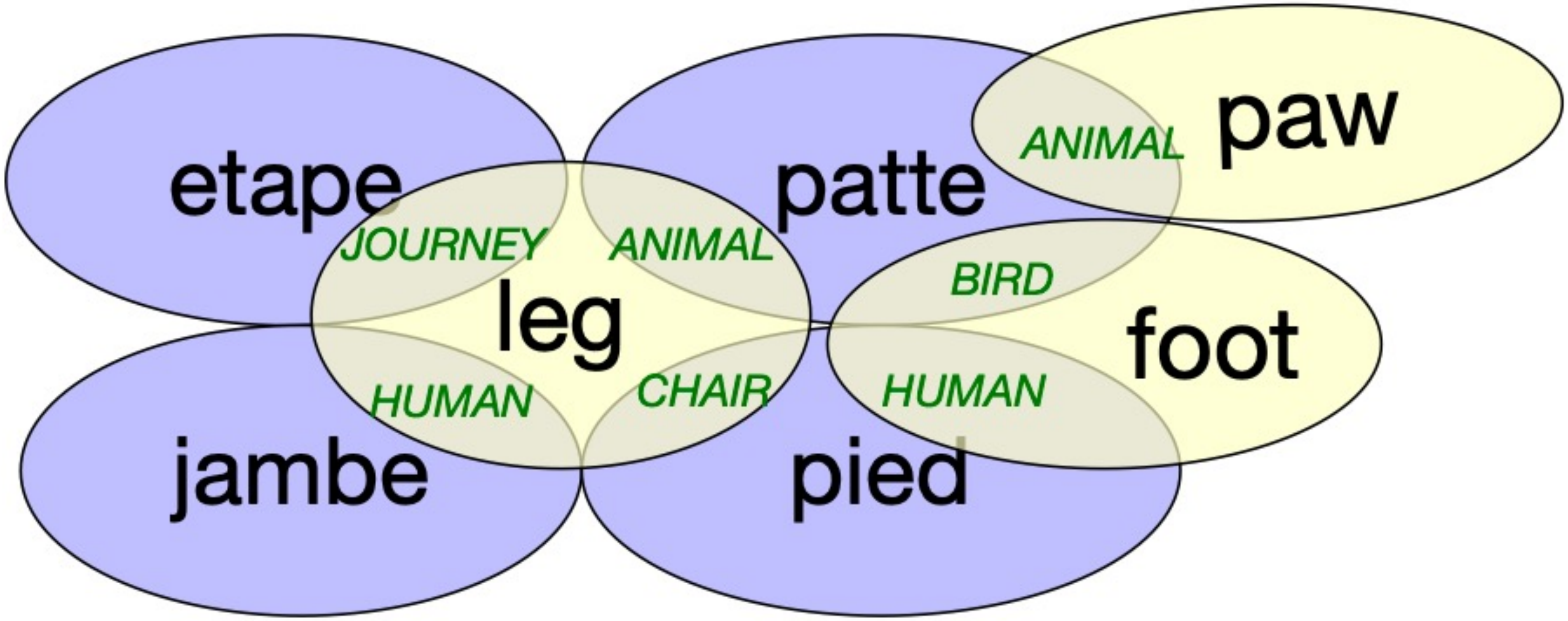


English:	*He adores listening to music*								
Japanese:	*kare*	*ha*	*ongaku*	*wo*	*kiku*	*no*	*ga*	*daisuki*	*desu*
	he		music	to	listening			adores	

  - Lexical level: words, alphabets are different.
  - Agglutination, ….

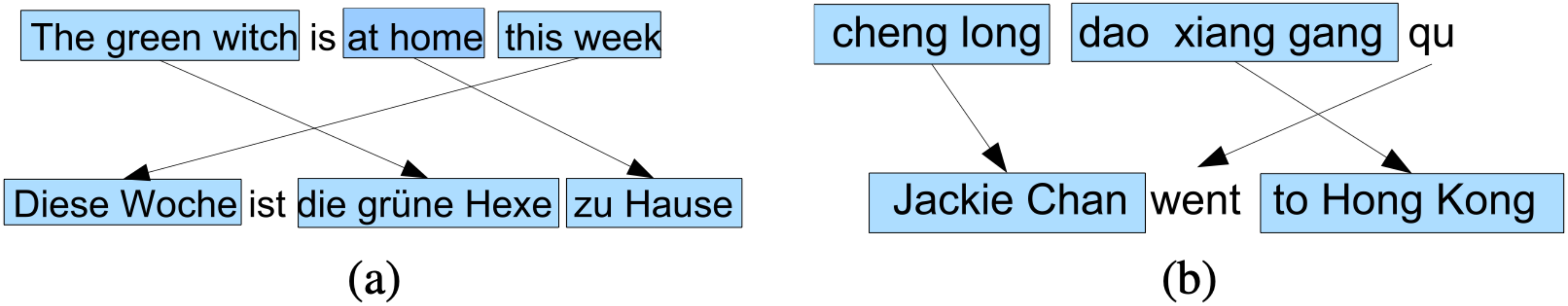
Why machine translation is hard?

The complex overlap between English leg, foot, etc. and various French translations like patte.



Why machine translation is hard?

- Complex reorderings may be needed.
- German often puts adverbs in initial position that English would put later.
- German tensed verbs often occur in second position causing the subject and verb to be inverted.



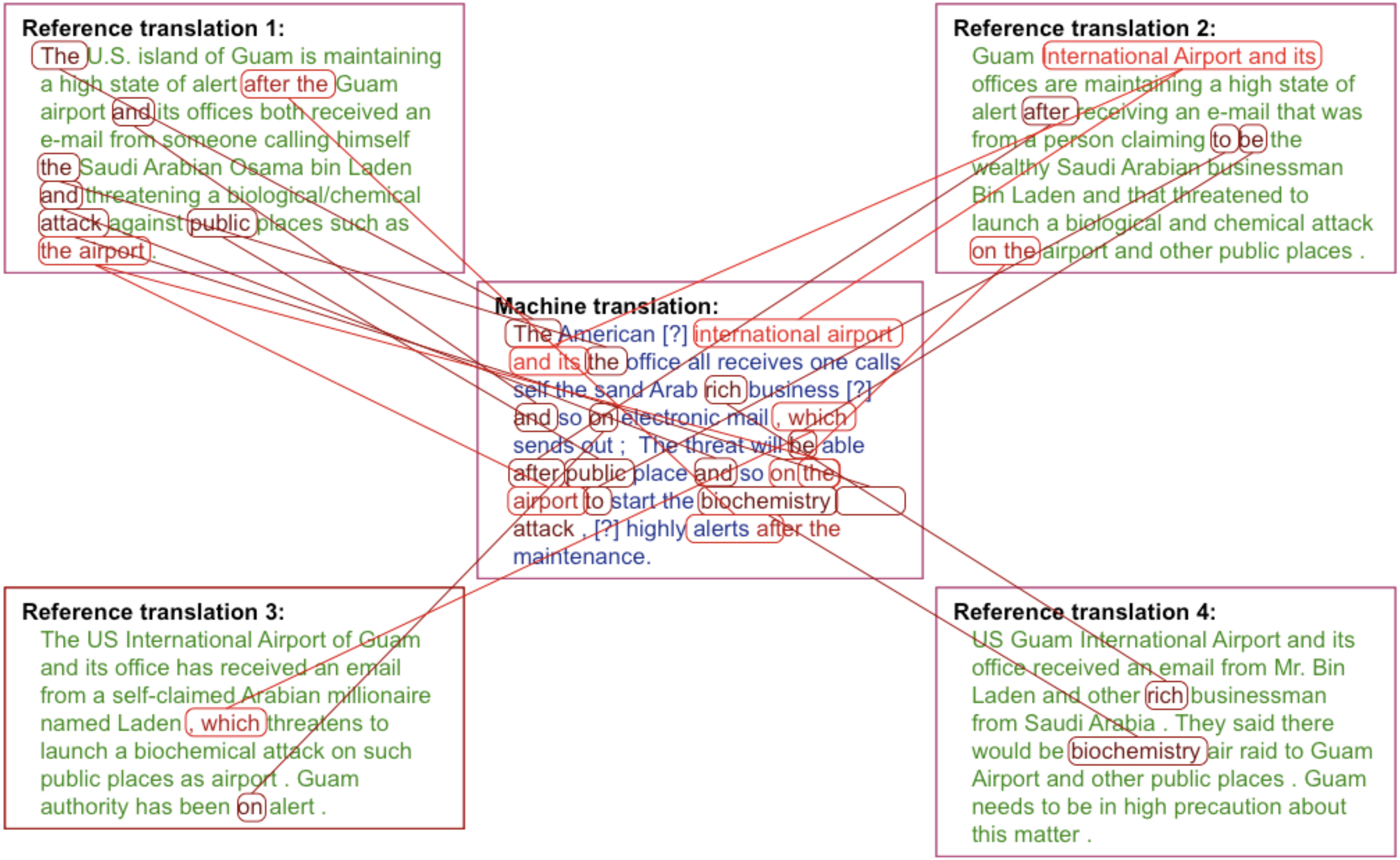
MT Evaluation

- Manual:
  - SSER (subjective sentence error rate)
  - Correct/Incorrect
  - Error categorization

- Testing in an application that uses MT as one sub-component
  - Question answering from foreign language documents
  - Cross Language Information Retrieval

- Automatic:
  - WER (word error rate)
  - **BLEU (Bilingual Evaluation Understudy)**
    - Proposed by IBM in 2001

Multiple reference translations



Reference translation 1:
The U.S. island of Guam is maintaining a high state of alert after the Guam airport and its offices both received an e-mail from someone calling himself the Saudi Arabian Osama bin Laden and threatening a biological/chemical attack against public places such as the airport .

Reference translation 2:
Guam International Airport and its offices are maintaining a high state of alert after receiving an e-mail that was from a person claiming to be the wealthy Saudi Arabian businessman Bin Laden and that threatened to launch a biological and chemical attack on the airport and other public places .

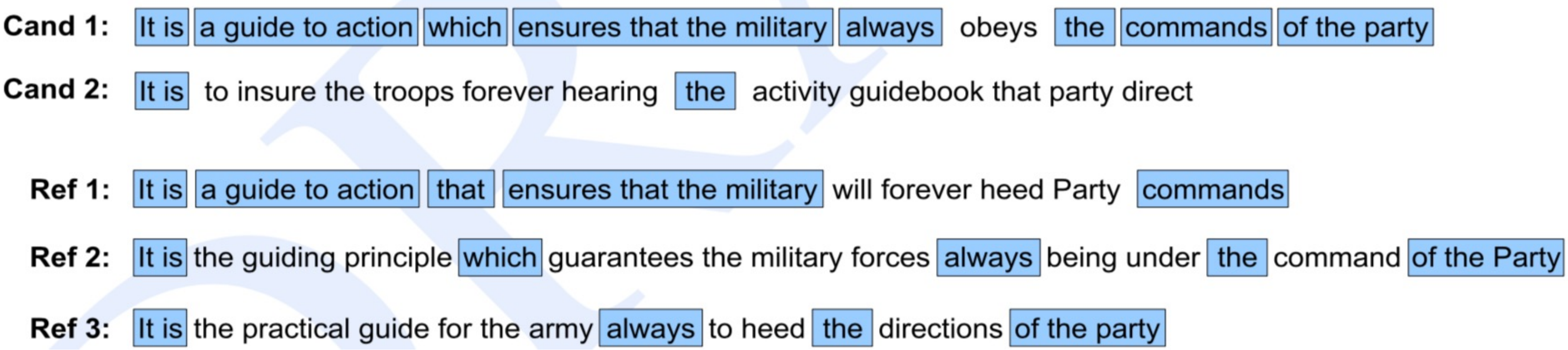
Machine translation:
The American [?] international airport and its the office all receives one calls self the sand Arab rich business [?] and so on electronic mail , which sends out ; The threat will be able after public place and so on the airport to start the biochemistry attack , [?] highly alerts after the maintenance.

Reference translation 3:
The US International Airport of Guam and its office has received an email from a self-claimed Arabian millionaire named Laden , which threatens to launch a biochemical attack on such public places as airport . Guam authority has been on alert .

Reference translation 4:
US Guam International Airport and its office received an email from Mr. Bin Laden and other rich businessman from Saudi Arabia . They said there would be biochemistry air raid to Guam Airport and other public places . Guam needs to be in high precaution about this matter .

BLEU Evaluation Metric

Intuition for BLEU: one of two candidate translations of a Chinese source sentence shares more words with the reference human translations.



Cand 1: It is a guide to action which ensures that the military always obeys the commands of the party

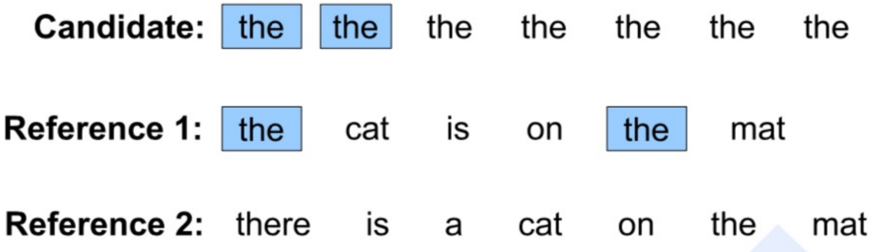
Cand 2: It is to insure the troops forever hearing the activity guidebook that party direct

Ref 1: It is a guide to action that ensures that the military will forever heed Party commands

Ref 2: It is the guiding principle which guarantees the military forces always being under the command of the Party

Ref 3: It is the practical guide for the army always to heed the directions of the party

A pathological example showing why BLEU uses a modified precision metric. Unigram precision would be unreasonably high (7/7). Modified unigram precision is appropriately low (2/7).



Candidate: the the the the the the the

Reference 1: the cat is on the mat

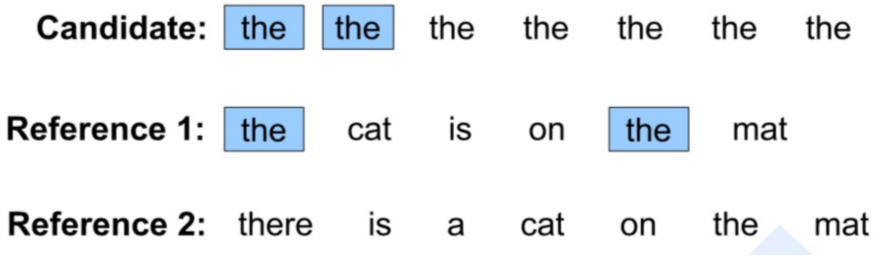
Reference 2: there is a cat on the mat

BLEU Evaluation Metric

- **Count** the maximum number of times a word is used in any single reference translation.
- The **Count** of each candidate word is then clipped by this maximum reference count.
- The modified precision score (for an entire test dataset):

$$p_n = \frac{\sum_{C \in \{Candidates\}} \sum_{n\text{-}gram \in C} \text{Count}_{\text{clip}}(n\text{-}gram)}{\sum_{C' \in \{Candidates\}} \sum_{n\text{-}gram' \in C'} \text{Count}(n\text{-}gram')}$$

- Here precision of 'the' is 2/7 not 7/7!



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

BLEU Evaluation Metric

Evaluation score:

$$\mathrm{Bleu} = \mathrm{BP} \times \exp\left(\frac{1}{N}\sum_{n=1}^{N} \log p_n\right)$$

Where BP is brevity penalty:

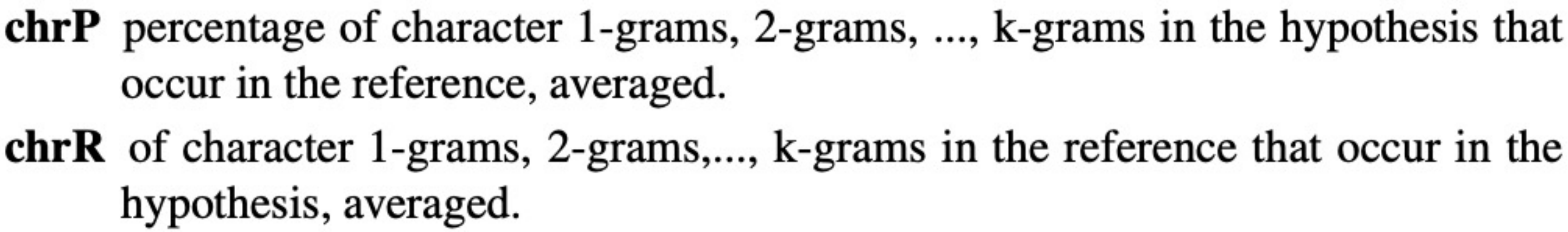
$$BP = \begin{cases} 1 & \text{if } \; c > r \\ e^{(1-r/c)} & \text{if } \; c \leq r \end{cases}$$

And p is normalized precision:

$$p_n = \frac{\displaystyle\sum_{C \in \{Candidates\}} \sum_{n\text{-}gram \in C} \mathrm{Count}_{\mathrm{clip}}(n\text{-}gram)}{\displaystyle\sum_{C' \in \{Candidates\}} \sum_{n\text{-}gram' \in C'} \mathrm{Count}(n\text{-}gram')}$$

chrF: character F-score

- Precision and Recall definition



chrP percentage of character 1-grams, 2-grams, ..., k-grams in the hypothesis that occur in the reference, averaged.

chrR of character 1-grams, 2-grams,..., k-grams in the reference that occur in the hypothesis, averaged.

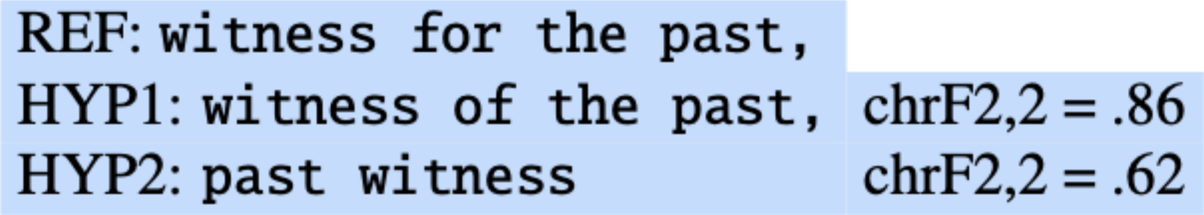
- Definition of the metric:

$$\text{chrF}\beta = (1+\beta^2)\frac{\text{chrP}\cdot\text{chrR}}{\beta^2\cdot\text{chrP}+\text{chrR}}$$

- Its commont to use F2 (beta = 2):

$$\text{chrF2} = \frac{5\cdot\text{chrP}\cdot\text{chrR}}{4\cdot\text{chrP}+\text{chrR}}$$

chrF: character F-score



REF: witness for the past,		
HYP1: witness of the past,	chrF2,2 = .86
HYP2: past witness	chrF2,2 = .62

unigrams that match: w i t n e s s f o t h e p a s t , (17 unigrams)

bigrams that match: wi it tn ne es ss th he ep pa as st t, (13 bigrams)

unigram P:	17/17 = 1	unigram R:	17/18 = .944
bigram P:	13/16 = .813	bigram R:	13/17 = .765

$$\mathrm{chrP} = (17/17 + 13/16)/2 = .906$$

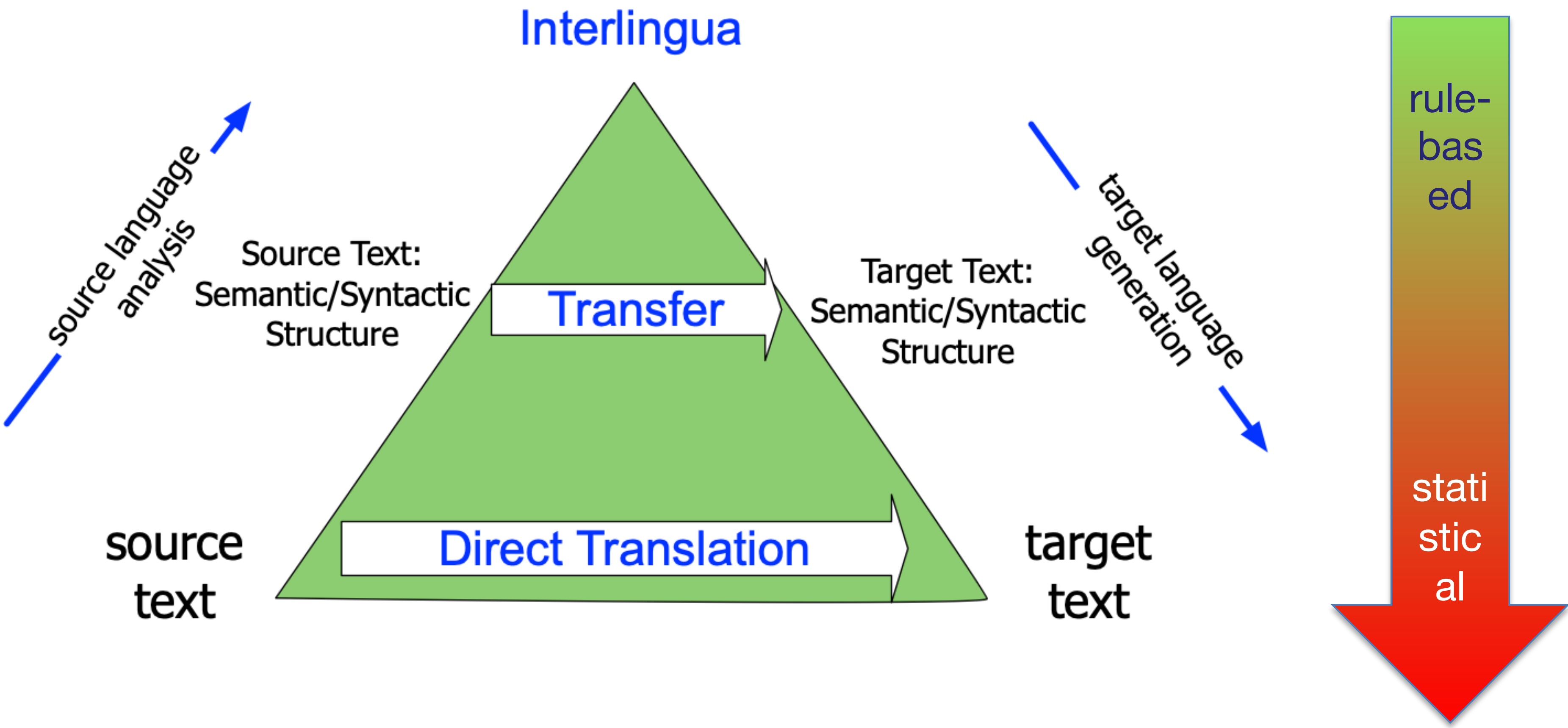
$$\mathrm{chrR} = (17/18 + 13/17)/2 = .855$$

$$\mathrm{chrF2,2} = 5\frac{\mathrm{chrP} * \mathrm{chrR}}{4\mathrm{chrP} + \mathrm{chrR}} = .86$$

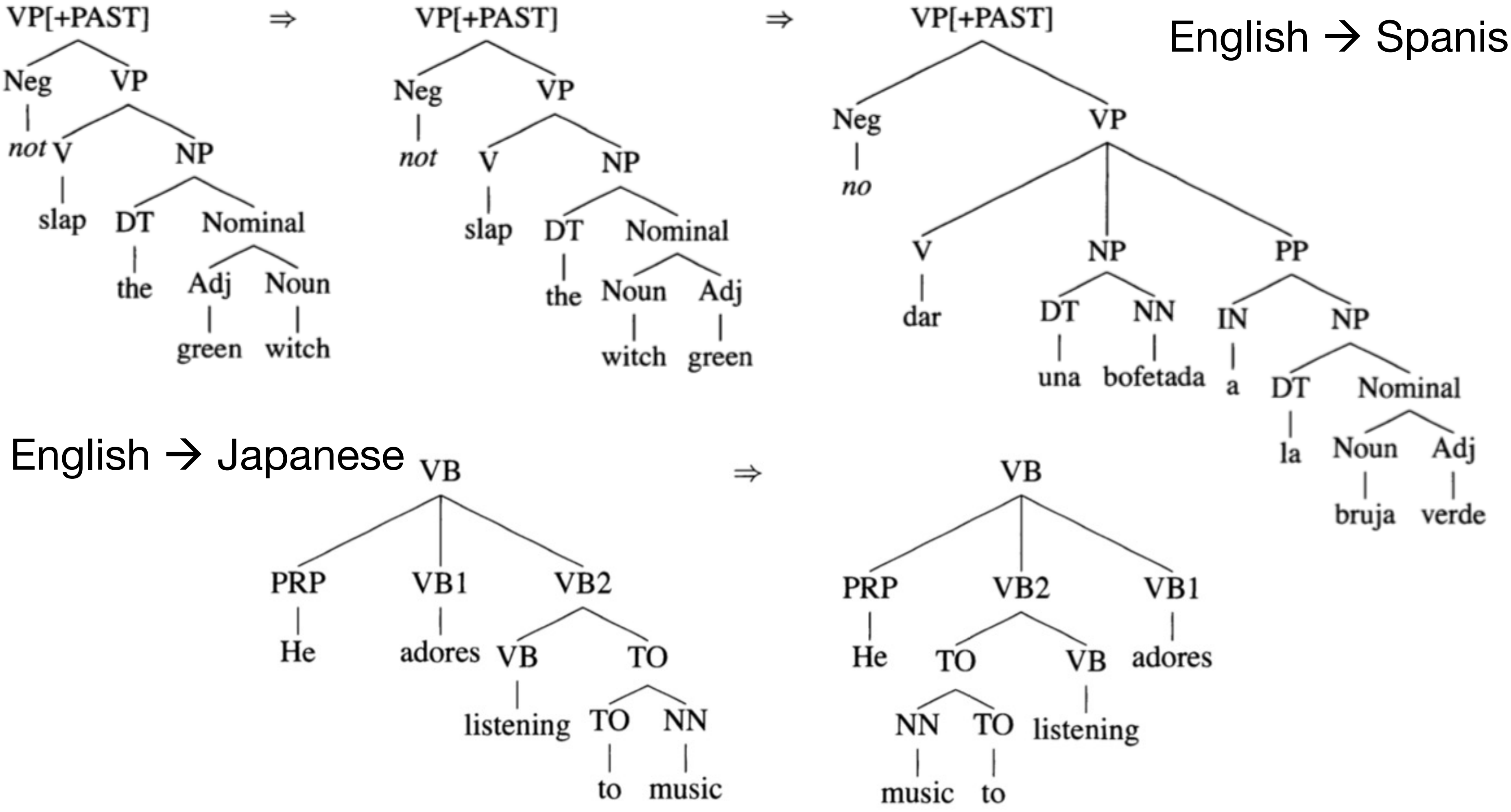
PLAN OF THE LECTURE

- Machine Translation (MT) Task
- Rule-based MT Models
- Statistical MT Models
- Neural MT Models

The Machine Translation Pyramid


Interlingua
source language analysis
Source Text: Semantic/Syntactic Structure
Transfer
Target Text: Semantic/Syntactic Structure
target language generation
source text
Direct Translation
target text
rule-bas ed
stati stic al

Rule-based statistical transfer

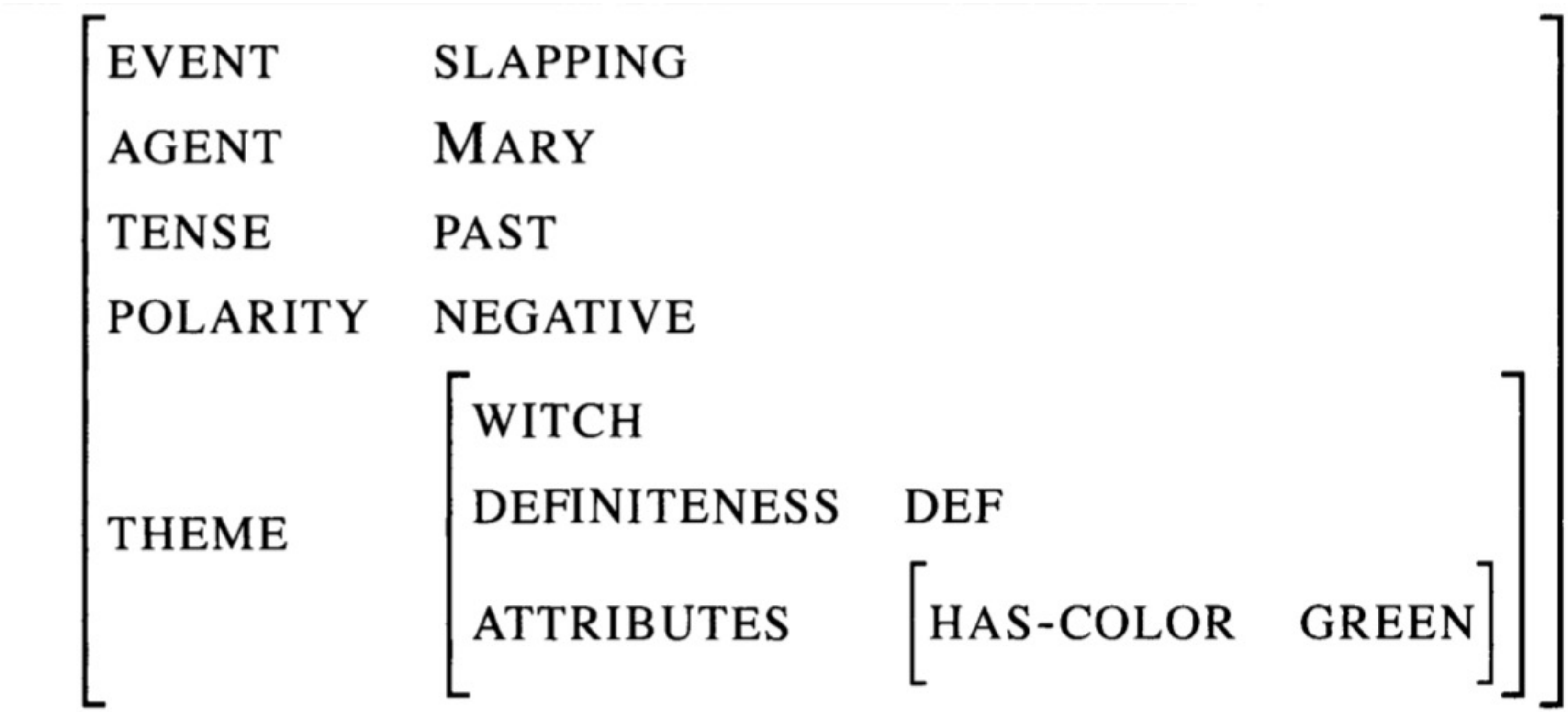


English → Spanish

English → Japanese

Interlingua

Interlingual representation of “Mary did not slap the green witch”.



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

PLAN OF THE LECTURE

- Machine Translation (MT) Task
- Rule-based MT Models
- Statistical MT Models
- Neural MT Models

A naïve approach: directly computing translation probabilities

Imagine that we want to translate from French (f) into English (e).

- Given a parallel corpus we can estimate *P*(e|f). The maximum likelihood estimation of *P*(e|f) is: *freq*(e,f)/*freq*(f)

A naïve approach: directly computing translation probabilities

Imagine that we want to translate from French (f) into English (e).

- Given a parallel corpus we can estimate *P*(e|f). The maximum likelihood estimation of *P*(e|f) is: *freq*(e,f)/*freq*(f)
- Way too specific to get any reasonable frequencies when done on the basis of sentences, vast majority of unseen data will have zero counts

A naïve approach: directly computing translation probabilities

Imagine that we want to translate from French (f) into English (e).

- Given a parallel corpus we can estimate *P*(e|f). The maximum likelihood estimation of *P*(e|f) is: *freq*(e,f)/*freq*(f)
- Way too specific to get any reasonable frequencies when done on the basis of sentences, vast majority of unseen data will have zero counts
- *P*(e|f) could be re-defined as:

$$P(e \mid f) = \prod_{f^j} \max_{e^i} P(e^i \mid f^j)$$

- **Problem**: The English words maximizing *P*(e|f) might not result in a readable sentence

Core idea of the statistical machine translation (SMT)

- We can account for adequacy: each foreign word translates into its most likely English word
- We cannot guarantee that this will result in a fluent English sentence

Core idea of the statistical machine translation (SMT): F → E

- We can account for adequacy: each foreign (F) word translates into its most likely English (E) word
- We cannot guarantee that this will result in a fluent English sentence
- **Solution**: transform P(E|F) with Bayes’ rule
- P(F|E) accounts for adequacy
- P(E) accounts for fluency

$$\hat{E} = \underset{E \in \text{English}}{\operatorname{argmax}} \overbrace{P(F|E)}^{\text{translation model}} \overbrace{P(E)}^{\text{language model}}$$

Best translation = faithfulness * fluency

$$\text{argmax}_y P(y|x) = \text{argmax}_y \underbrace{P(x|y)}_{} \underbrace{P(y)}_{}$$

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.

http://web.stanford.edu/class/cs224n/slides/cs224n-2020-lecture08-nmt.pdf

Three Tasks of Statistical MT

- Language model
  - Given a target language string e, assigns P(e)
  - good target language string ➔ high P(e)
  - random word sequence ➔ low P(e)

- Translation model
  - Given a pair of strings <f,e>, assigns P(f|e) by formula
  - <f,e> look like translations ➔ high P(f|e)
  - <f,e> don't look like translations ➔ low P(f|e)

- Decoding algorithm
  - Given a language model, a translation model, and a new sentence f: find translation e maximizing P(e)•P(f|e)

Language Modeling: P(e)

- Determine the probability of an English sequence P(e)
- Can use n-gram models, PCFG-based models etc.: anything that assigns a probability for a sequence
- Standard: n-gram model

$$P(e) = P(e^1)P(e^2 \mid e^1)\prod_{i=3}^{l} P(e^i \mid e^{i-1}..e^{i-n+1})$$

Language Modeling: P(e)

- Determine the probability of an English sequence P(e)
- Can use n-gram models, PCFG-based models etc.: anything that assigns a probability for a sequence
- Standard: n-gram model

$$P(e) = P(e^1)P(e^2 \mid e^1)\prod_{i=3}^{l} P(e^i \mid e^{i-1}..e^{i-n+1})$$

- Language model picks the most fluent translation of many possible translations
- Language model can be estimated from a large monolingual corpus

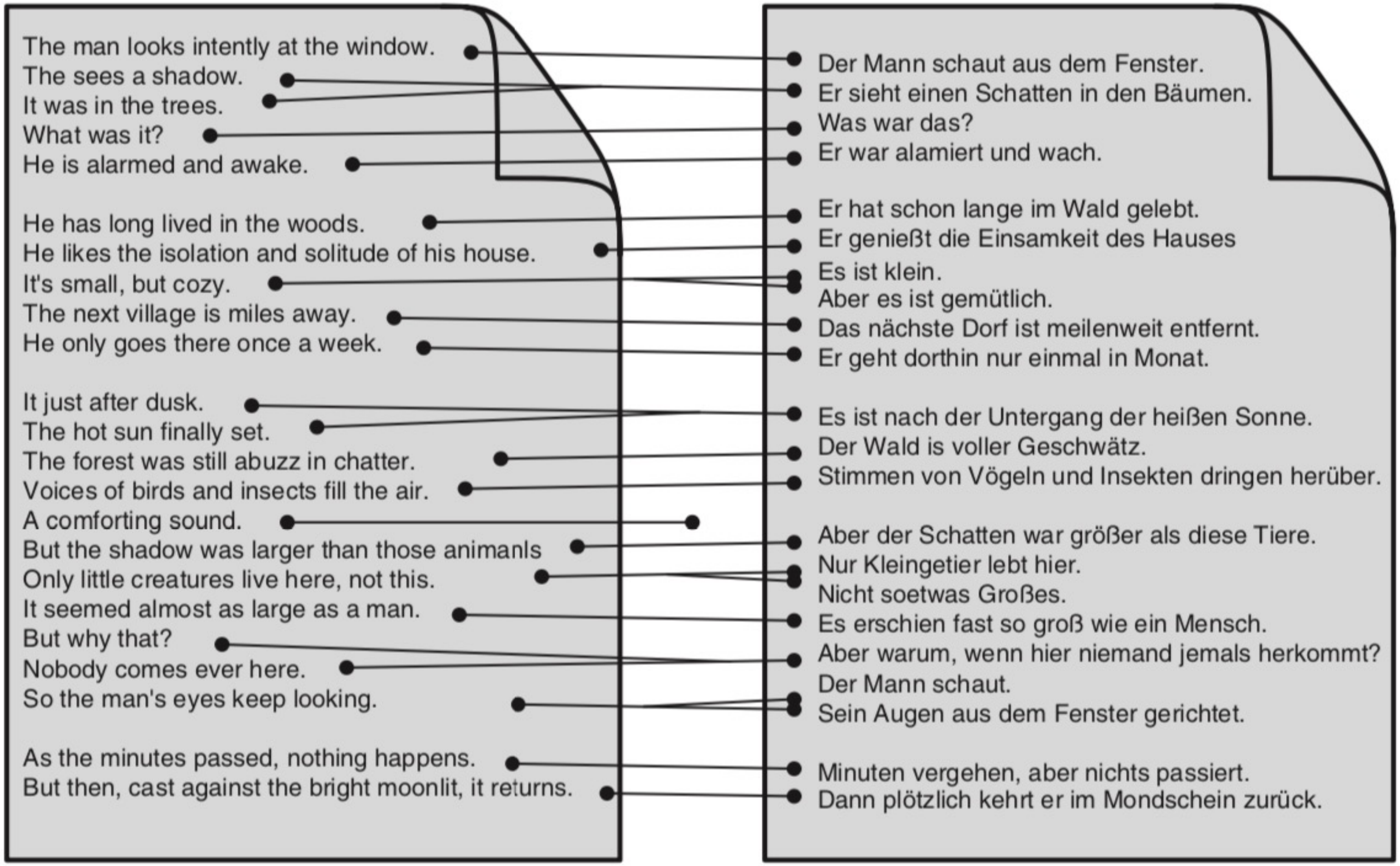
Translation Modeling: P(f|e)

- Determines the probability that the foreign word f_j is a translation of the English word e_i
- How to compute $P(f_j \mid e_i)$ from a parallel corpus? Need to **align** their translations
- Statistical approaches rely on the co-occurrence of e_i and f_j in the parallel data: If e_i and f_j tend to co-occur in parallel sentence pairs, they are likely to be translations of one another

Translation Modeling: P(f|e)

- Determines the probability that the foreign word f_j is a translation of the English word e_i
- How to compute $P(f_j \mid e_i)$ from a parallel corpus? Need to **align** their translations
- Statistical approaches rely on the co-occurrence of e_i and f_j in the parallel data: If e_i and f_j tend to co-occur in parallel sentence pairs, they are likely to be translations of one another
- Commonly, four factors are used:
  - **translation:** How often do e_i and f_j co-occur?
  - **distortion:** How likely is a word occurring at position x to translate into a word occurring at position y? For example: English is a verb-second language, whereas German is a verb-final language
  - **fertility**: How likely is e_i to translate into more than one word? For example: “defeated” can translate into "eine Niederlage erleiden"
  - **null translation**: How likely is a foreign word to be spuriously generated?

Sentence alignment



The man looks intently at the window.
The sees a shadow.
It was in the trees.
What was it?
He is alarmed and awake.

He has long lived in the woods.
He likes the isolation and solitude of his house.
It's small, but cozy.
The next village is miles away.
He only goes there once a week.

It just after dusk.
The hot sun finally set.
The forest was still abuzz in chatter.
Voices of birds and insects fill the air.
A comforting sound.
But the shadow was larger than those animanls
Only little creatures live here, not this.
It seemed almost as large as a man.
But why that?
Nobody comes ever here.
So the man's eyes keep looking.

As the minutes passed, nothing happens.
But then, cast against the bright moonlit, it returns.

Der Mann schaut aus dem Fenster.
Er sieht einen Schatten in den Bäumen.
Was war das?
Er war alamiert und wach.

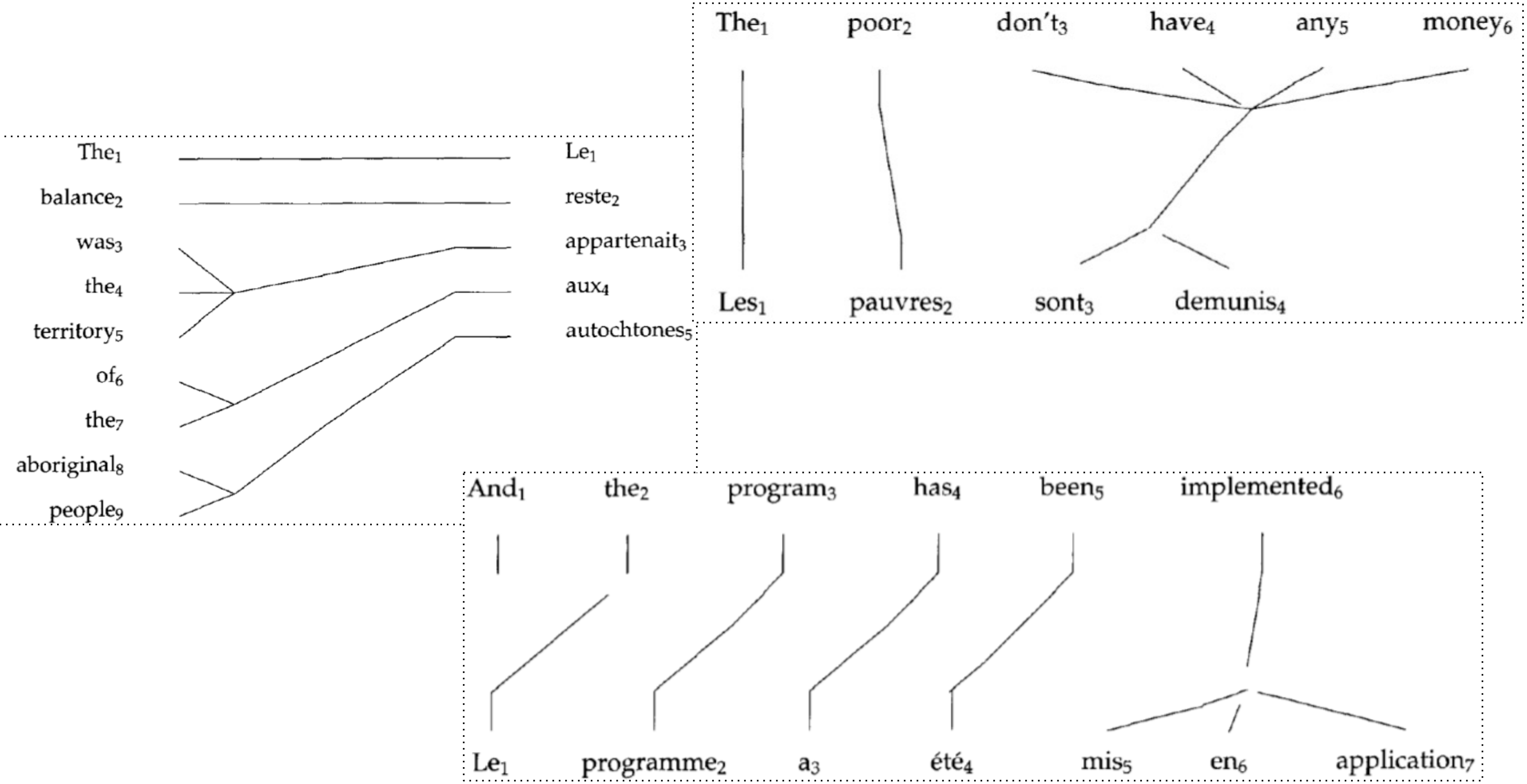
Er hat schon lange im Wald gelebt.
Er genießt die Einsamkeit des Hauses
Es ist klein.
Aber es ist gemütlich.
Das nächste Dorf ist meilenweit entfernt.
Er geht dorthin nur einmal in Monat.

Es ist nach der Untergang der heißen Sonne.
Der Wald is voller Geschwätz.
Stimmen von Vögeln und Insekten dringen herüber.

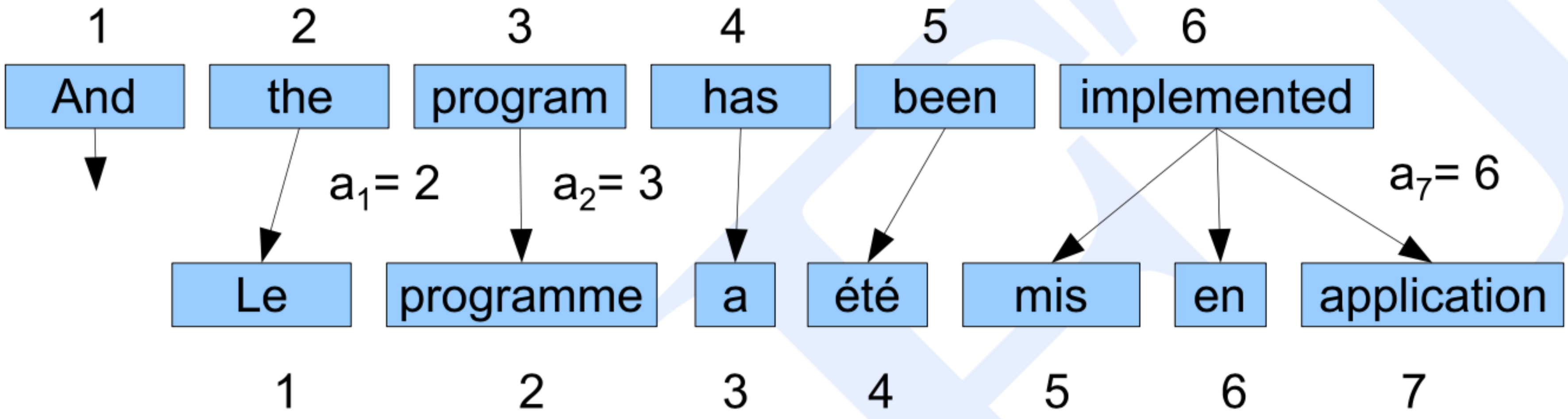
Aber der Schatten war größer als diese Tiere.
Nur Kleingetier lebt hier.
Nicht soetwas Großes.
Es erschien fast so groß wie ein Mensch.
Aber warum, wenn hier niemand jemals herkommt?
Der Mann schaut.
Sein Augen aus dem Fenster gerichtet.

Minuten vergehen, aber nichts passiert.
Dann plötzlich kehrt er im Mondschein zurück.

Word Alignment



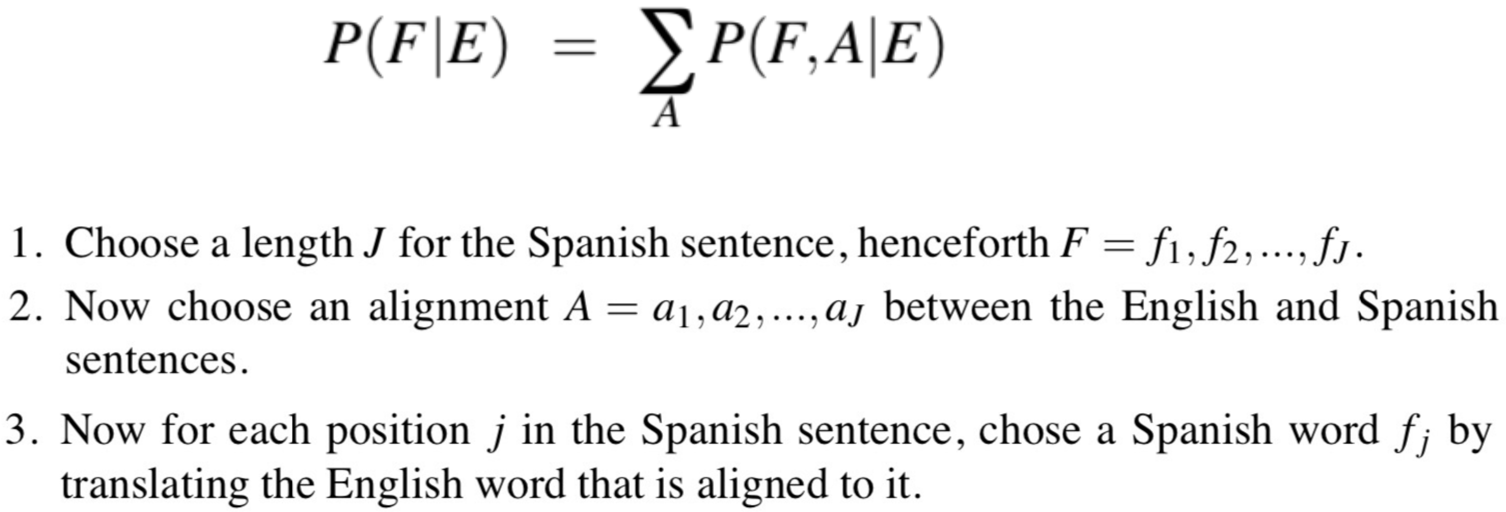
Word Alignment



A = 2, 3, 4, 5, 6, 6, 6

IBM Model 1

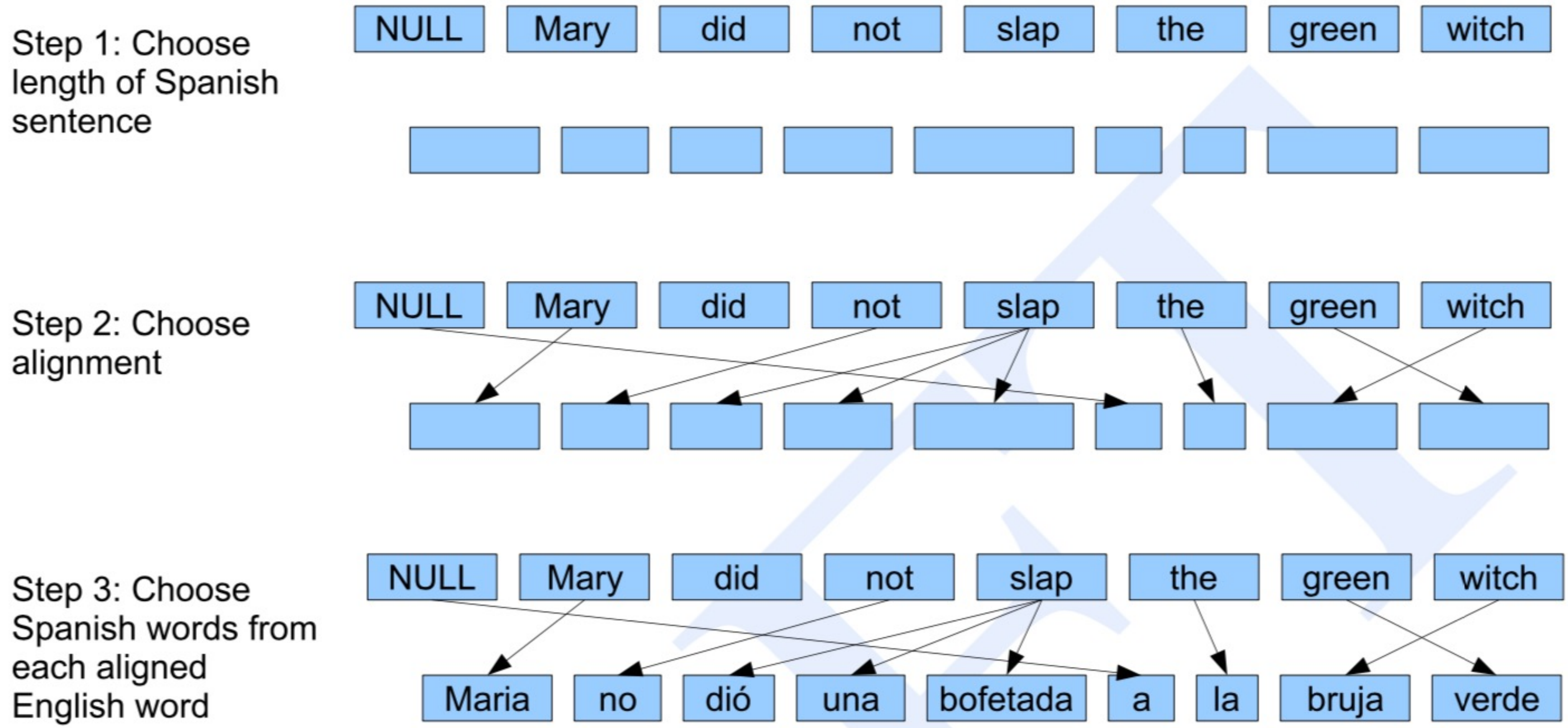
- Simplest of the IBM models
- Does not model one-to-many alignments
- Computationally inexpensive



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

1. Choose a length J for the Spanish sentence, henceforth $F = f_1, f_2, ..., f_J$.
2. Now choose an alignment $A = a_1, a_2, ..., a_J$ between the English and Spanish sentences.
3. Now for each position j in the Spanish sentence, chose a Spanish word f_j by translating the English word that is aligned to it.

IBM Model 1: generative story



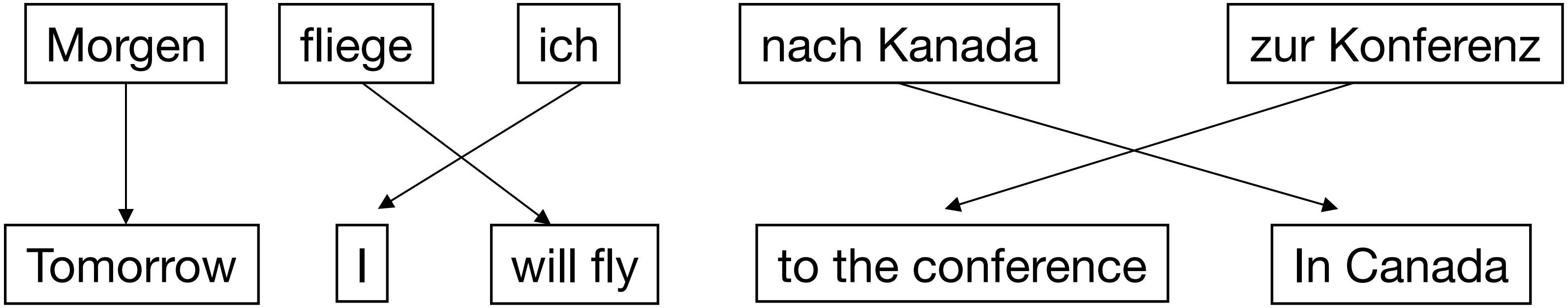
IBM Models by Brown et al. (1993)

- Model 1: lexical translation
  - Bag of words
  - Unique local maxima
  - Efficient EM algorithm
- Model 2: adds absolute alignment model:

  $$a(e^{pos} \mid f^{pos}, e_{length}, f_{length})$$
- Model 3: add fertility model: n(k|e)
  - No full EM, count only neighbors (Model 3–5)
  - Leaky (Model 3–4)
- Model 4: adds relative alignment model
  - Relative distortion
  - word classes
- Model 5: fixes deficiency
  - Extra variables to avoid leakiness

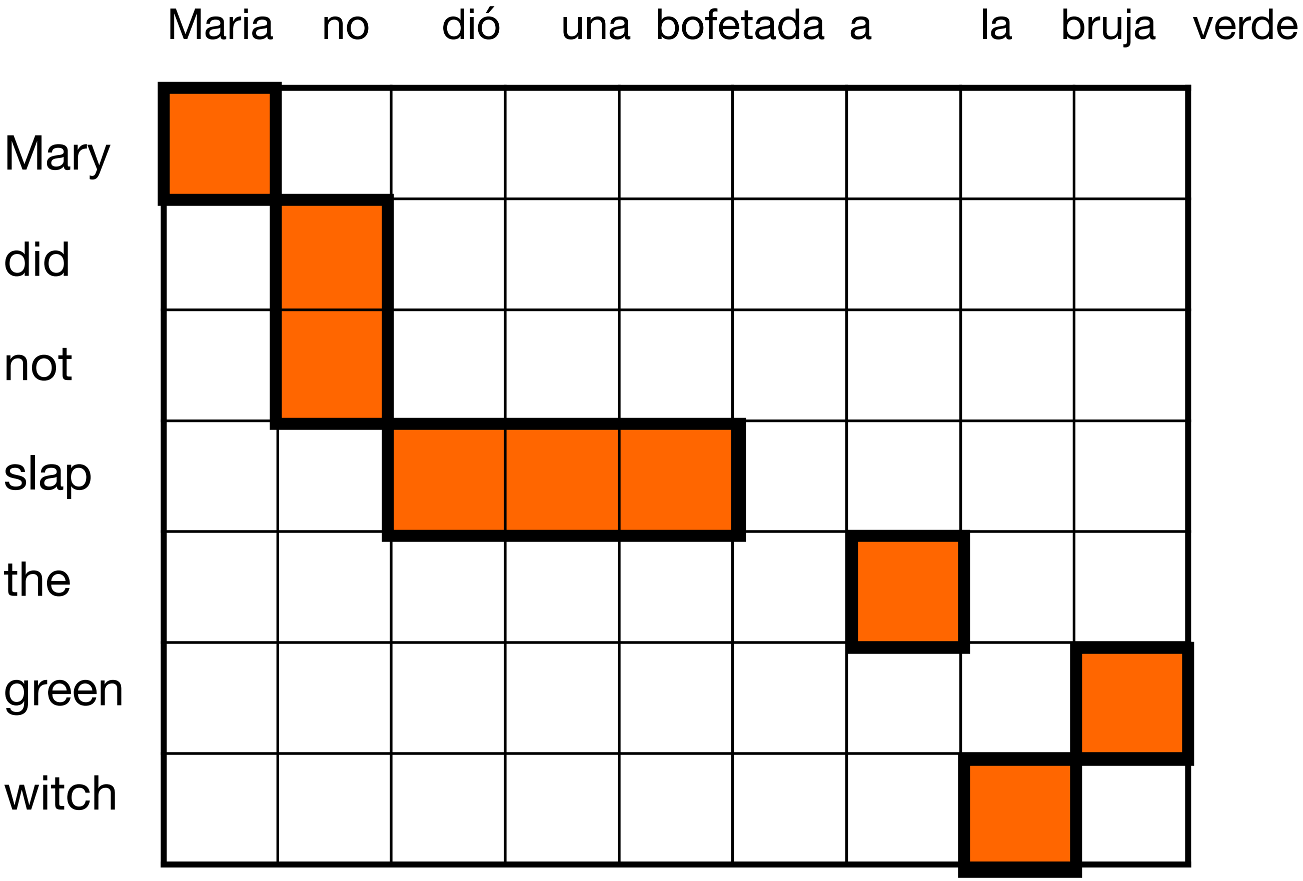
Brown, P. F., Pietra, V. J. D., Pietra, S. A. D., & Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. Computational linguistics, 19(2), 263-311.

Phrase-Based Statistical MT



- Foreign input segmented into phrases
- “phrase” is any sequence of words
- Each phrase is probabilistically translated into English
  - P(to the conference | zur Konferenz)
  - P(into the meeting | zur Konferenz)
- Phrases are probabilistically re-ordered

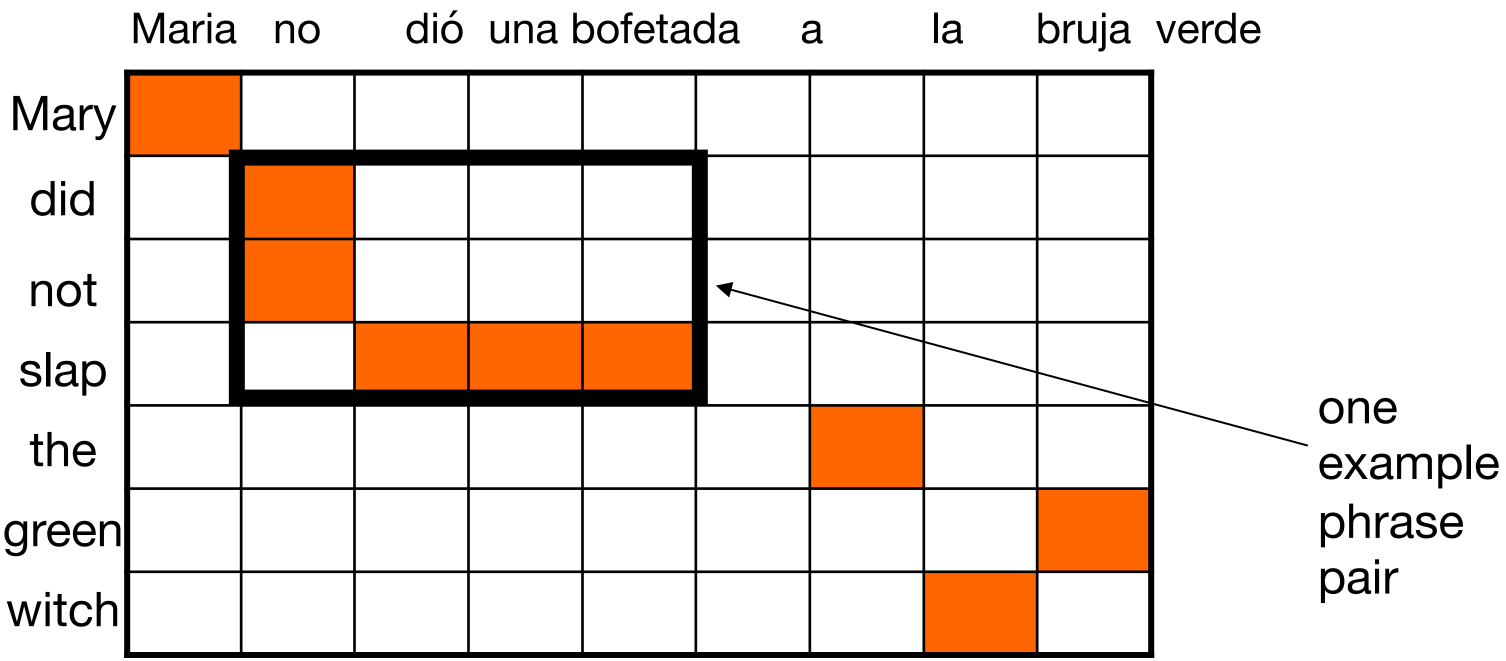
Word Alignment Induced Phrases



	Maria	no	dió	una	bofetada	a	la	bruja	verde
Mary	■								
did		■							
not		■							
slap			■	■	■				
the							■		
green									■
witch								■	

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

How to learn the phrase translation table?

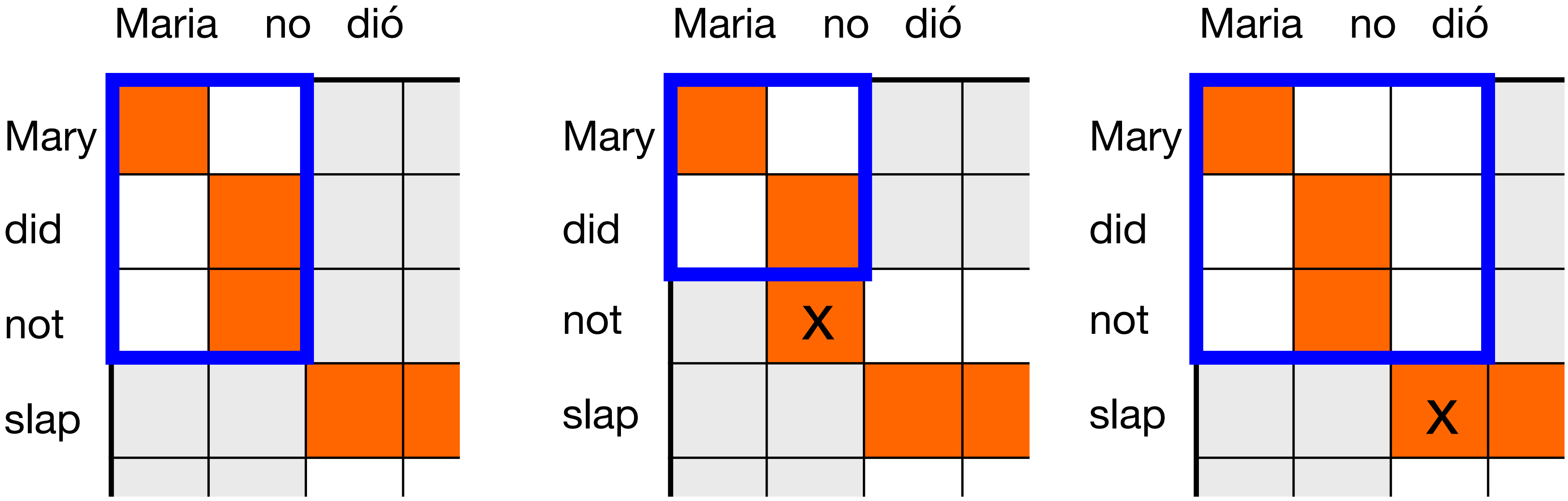


	Maria	no	dió	una	bofetada	a	la	bruja	verde
Mary	■								
did		■							
not		■							
slap			■	■	■				
the							■		
green									■
witch								■	

one example phrase pair

- Collect **all phrase pairs** that are **consistent** with the word alignment

Consistent with Word Alignment



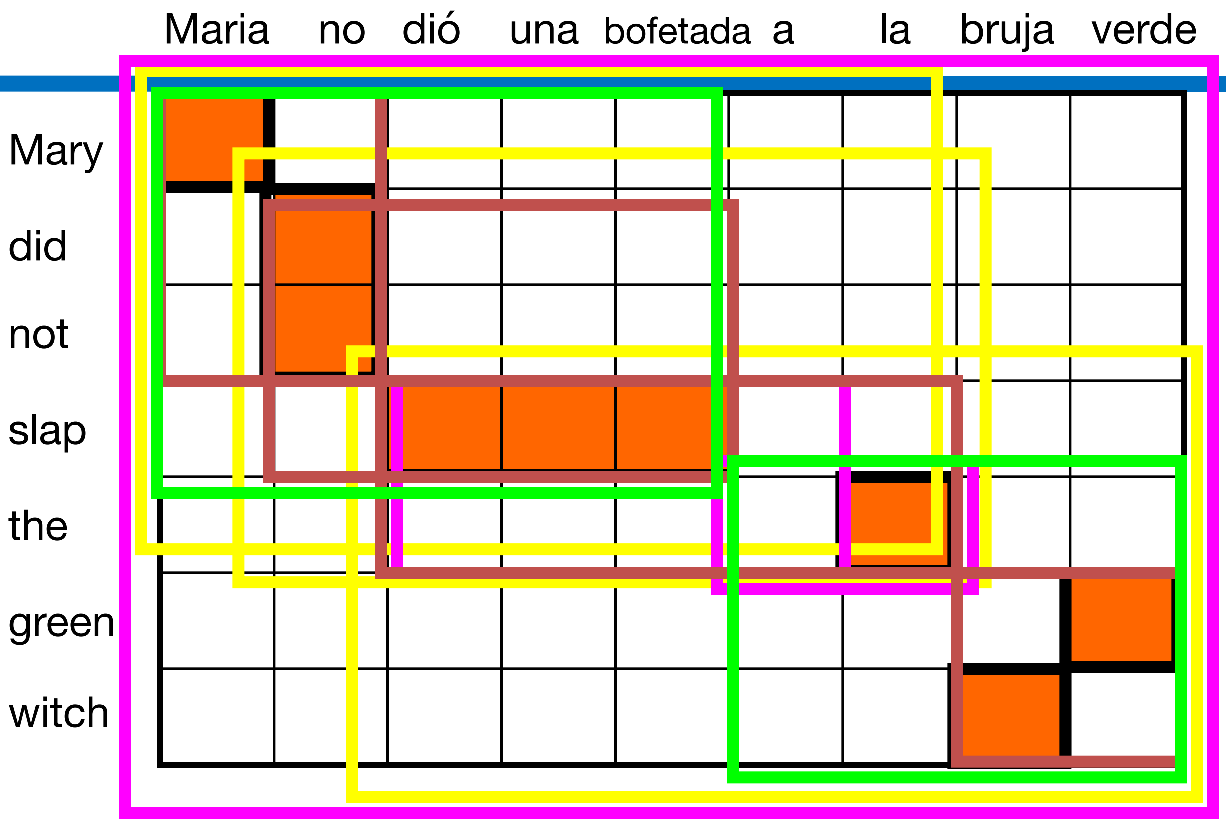
consistent

inconsistent

inconsistent

- Phrase alignment must contain all alignment points for all the words in both phrases!

Word Alignment Induced Phrases



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

Decoding

- Goal is to find a translation that maximizes the product of the translation and language models.

$$\underset{e}{\operatorname{argmax}} P(f \mid e) P(e)$$

- Cannot explicitly enumerate and test the combinatorial space of all possible translations.
- Must efficiently (heuristically) search the space of translations that approximates the solution to this difficult optimization problem.

Pharaoh: a Beam Search Decoder for Phrase-Based Statistical Machine Translation Models, Philipp Koehn, AMTA 2004

Decoding

- Goal is to find a translation that maximizes the product of the translation and language models.

$$\underset{e}{\text{argmax}}\, P(f \mid e)P(e)$$

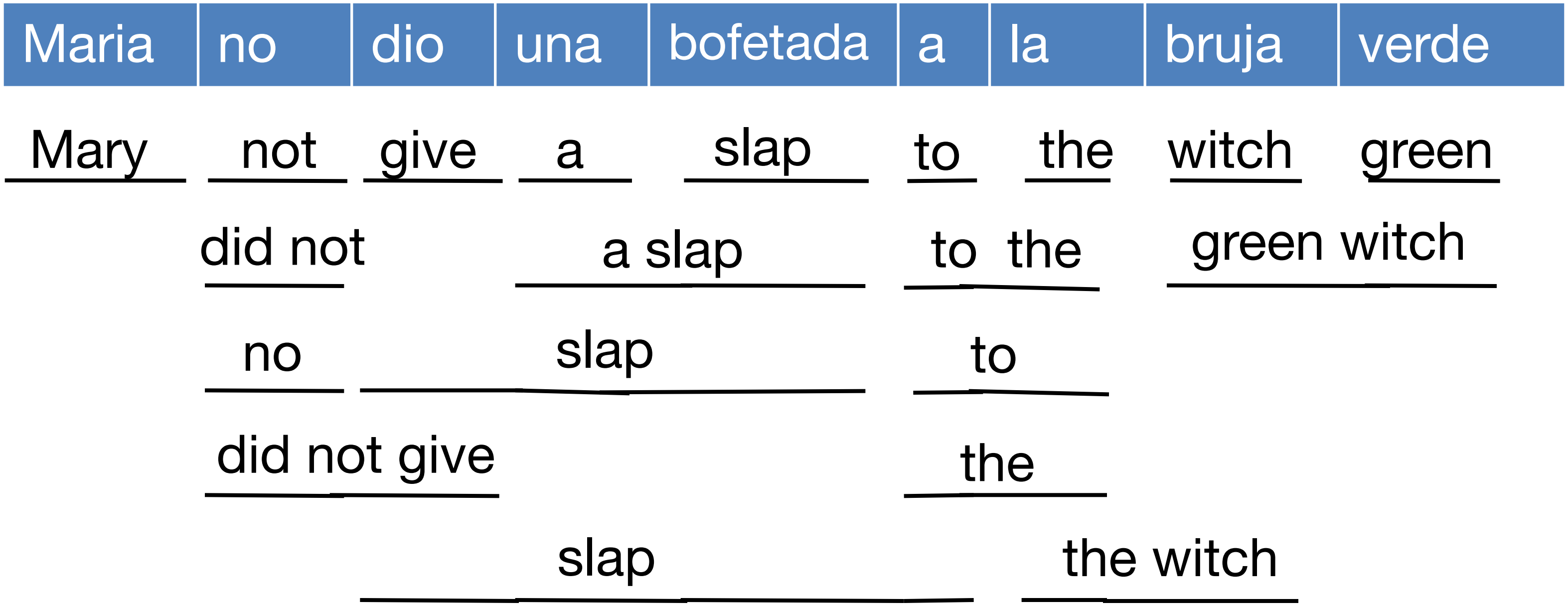
- Cannot explicitly enumerate and test the combinatorial space of all possible translations.
- Must efficiently (heuristically) search the space of translations that approximates the solution to this difficult optimization problem.
- The optimal decoding problem for all reasonable models (e.g. IBM model 1) is NP-complete.

Here:

- phrase-based decoder based on that of Koehn's (2004) Pharaoh system.

Pharaoh: a Beam Search Decoder for Phrase-Based Statistical Machine Translation Models, Philipp Koehn, AMTA 2004

Space of Translations



Maria	no	dio	una	bofetada	a	la	bruja	verde

Mary not give a slap to the witch green

did not a slap to the green witch

no slap to

did not give the

slap the witch

The phrase translation table from the alignment defines the space of possible translations

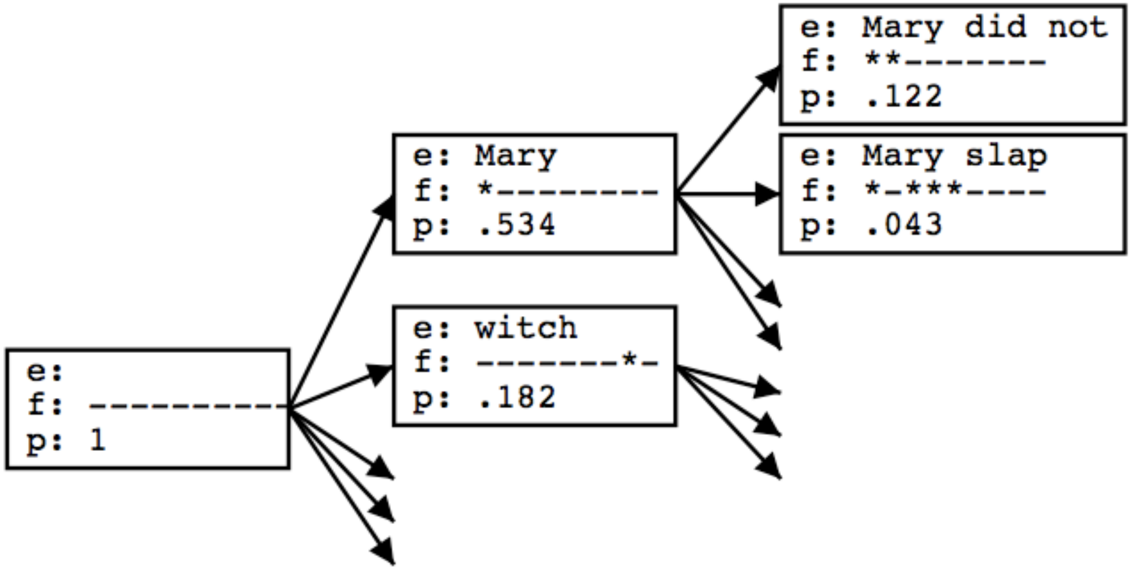
- every word can have multiple translations
- every word can participate in multiple phrases

Stack Decoding

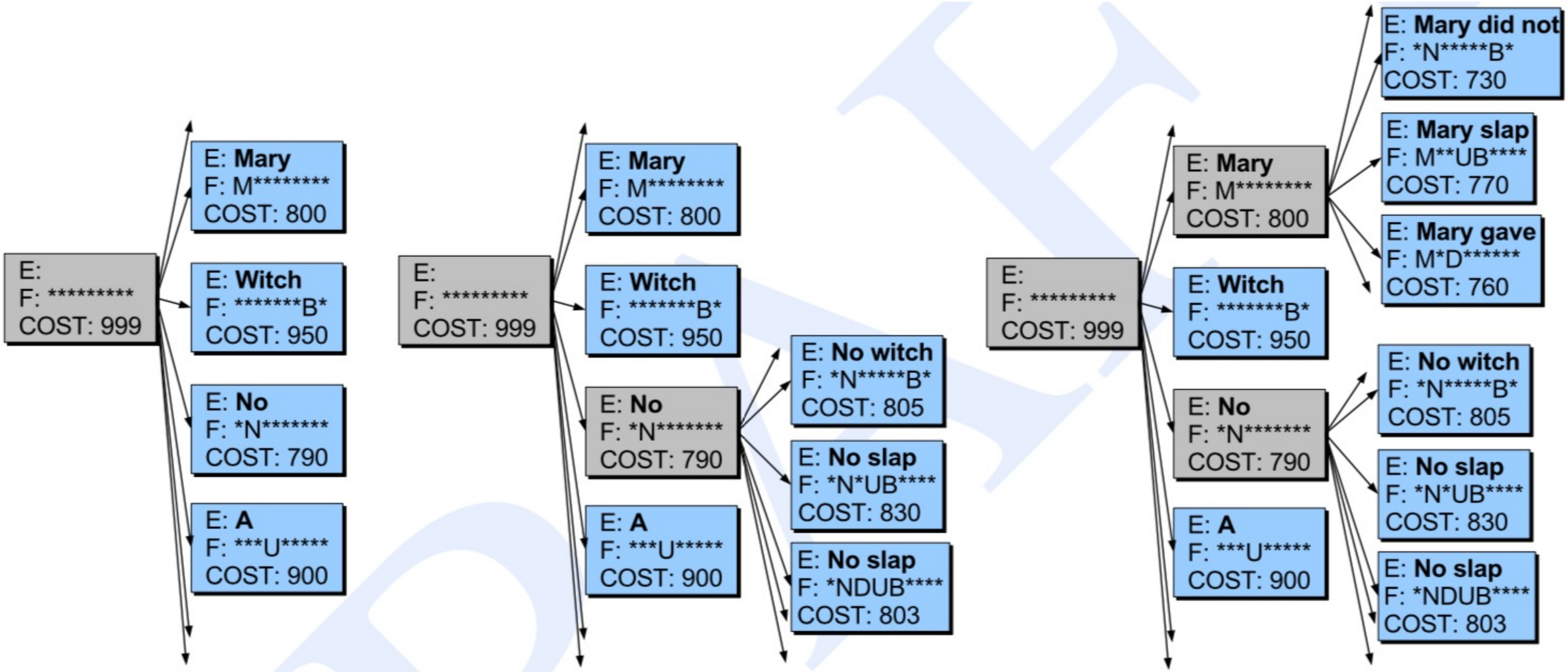
- Use a version of heuristic A* search to explore the space of phrase translations to find the best scoring subset that covers the source sentence.

```
function STACK DECODING(source sentence) returns target sentence

initialize stack with a null hypothesis
loop do
    pop best hypothesis $h$ off of stack
    if $h$ is a complete sentence, return $h$
    for each possible expansion $h'$ of $h$
        assign a score to $h'$
        push $h'$ onto stack
```



Stack Decoding



E:
F: *********
COST: 999

E: **Mary**
F: M********
COST: 800

E: **Witch**
F: *******B*
COST: 950

E: **No**
F: *N*******
COST: 790

E: **A**
F: ***U*****
COST: 900

a) after expanding NULL

E:
F: *********
COST: 999

E: **Mary**
F: M********
COST: 800

E: **Witch**
F: *******B*
COST: 950

E: **No**
F: *N*******
COST: 790

E: **A**
F: ***U*****
COST: 900

E: **No witch**
F: *N*****B*
COST: 805

E: **No slap**
F: *N*UB****
COST: 830

E: **No slap**
F: *NDUB****
COST: 803

b) after expanding "No"

E:
F: *********
COST: 999

E: **Mary**
F: M********
COST: 800

E: **Witch**
F: *******B*
COST: 950

E: **No**
F: *N*******
COST: 790

E: **A**
F: ***U*****
COST: 900

E: **Mary did not**
F: *N*****B*
COST: 730

E: **Mary slap**
F: M**UB****
COST: 770

E: **Mary gave**
F: M*D******
COST: 760

E: **No witch**
F: *N*****B*
COST: 805

E: **No slap**
F: *N*UB****
COST: 830

E: **No slap**
F: *NDUB****
COST: 803

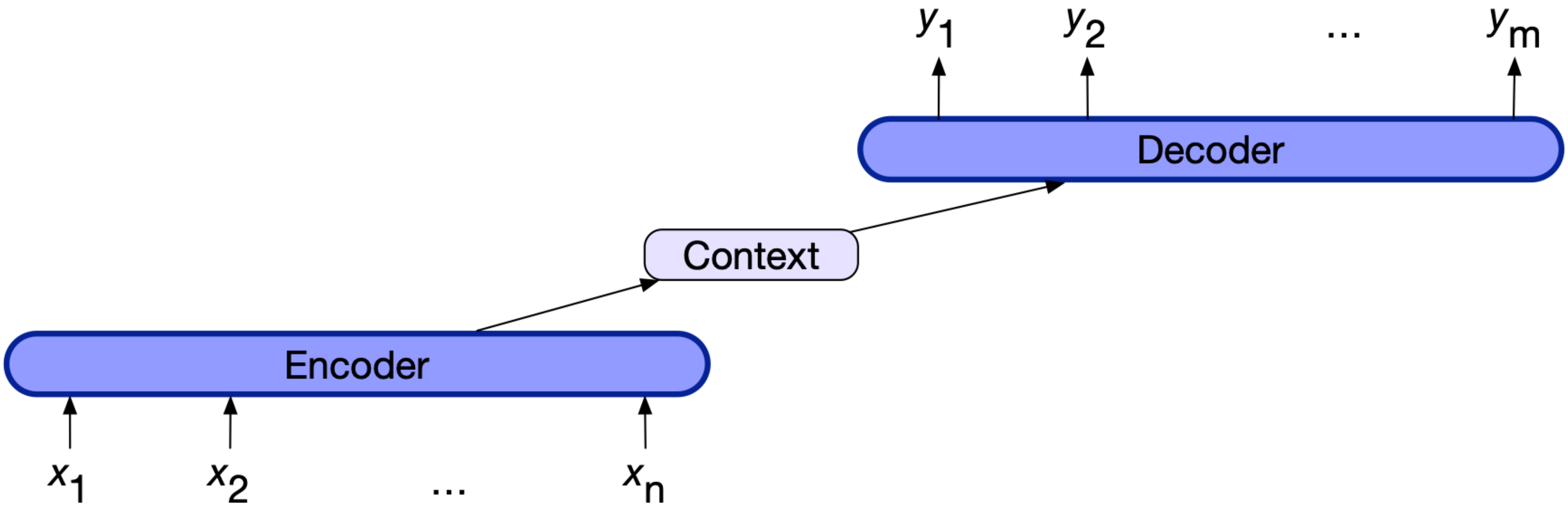
c) after expanding "Mary"

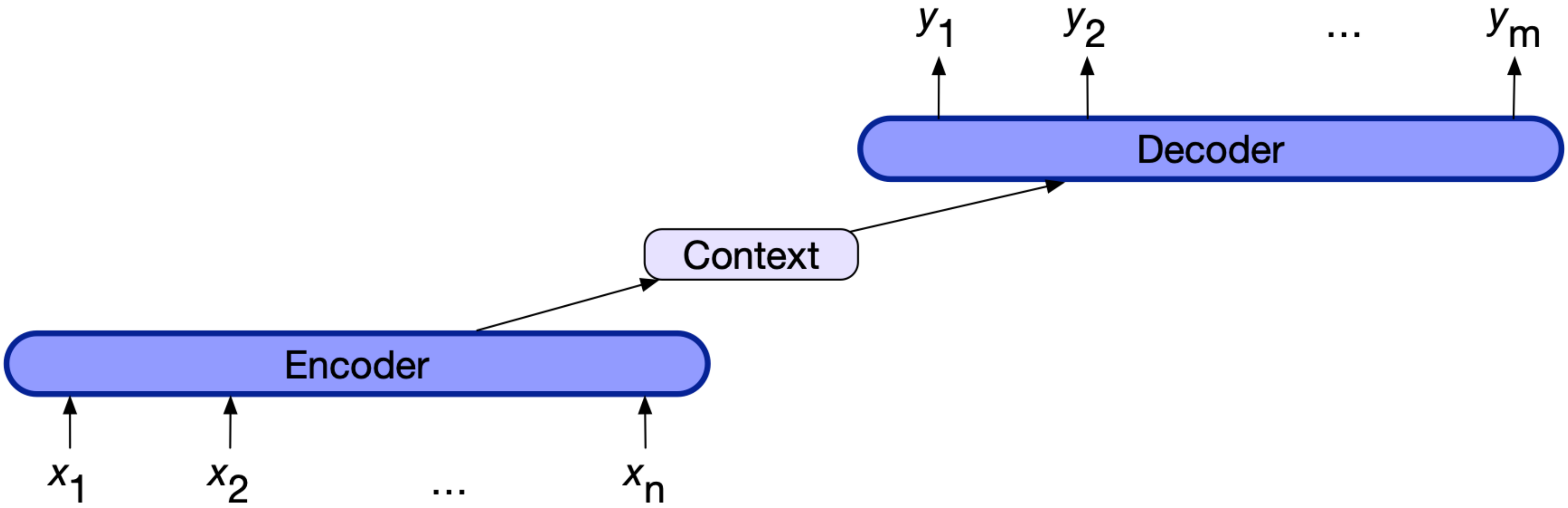
PLAN OF THE LECTURE

- Machine Translation (MT) Task
- Rule-based MT Models
- Statistical MT Models
- Neural MT Models

Encoder-Decoder Model

- Use of an **encoder network** that takes an input sequence and creates a contextualized representation of it, often called the **context.**
- This representation is then passed to a **decoder** which generates a task-specific output sequence.



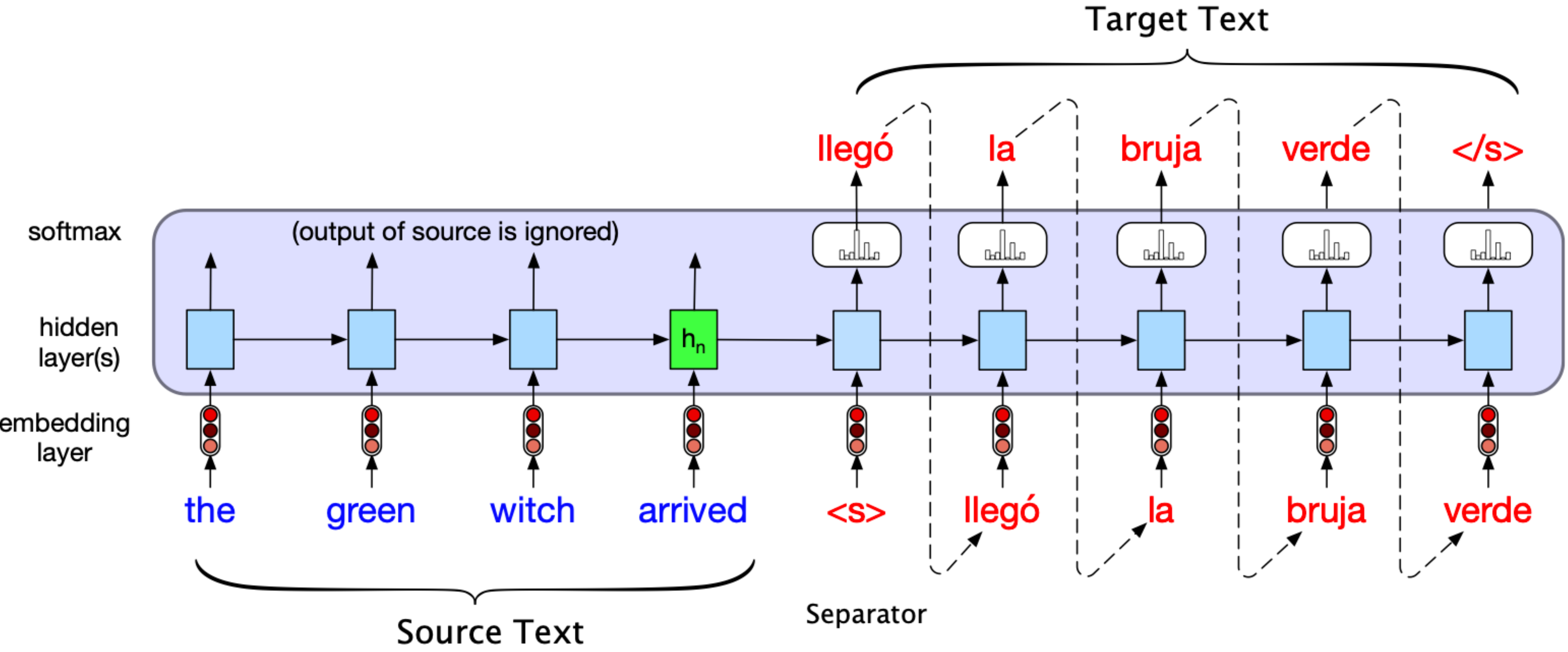


- An **encoder** that accepts an input sequence, and generates a corresponding sequence of contextualized representations.
  - LSTMs, GRUs, convolutional networks, and Transformers can all be employed as encoders
- A **context** vector is a function of hidden representation conveys the essence of the input to the decoder.
- A **decoder**, which accepts context vector as input and generates an arbitrary length sequence of hidden states, from which a corresponding sequence of output states, can be obtained.
  - Just as with encoders, decoders can be realized by any kind of sequence architecture.

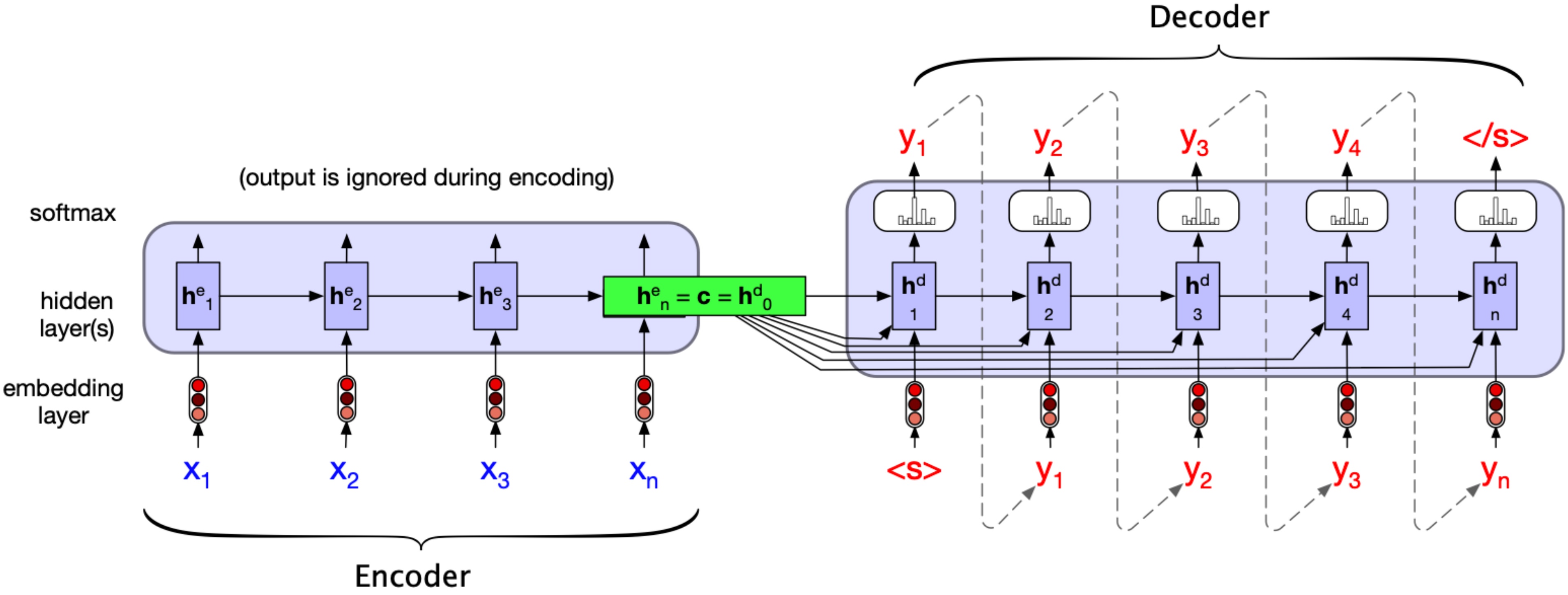
Encoder-Decoder with RNNs

- Given source text x and target text y:

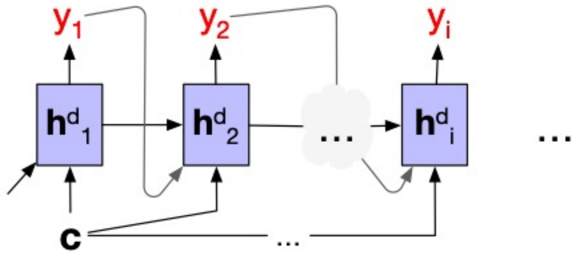
$$p(y|x) \;=\; p(y_1|x)p(y_2|y_1,x)p(y_3|y_1,y_2,x)...P(y_m|y_1,...,y_{m-1},x)$$



Encoder-Decoder with RNNs



Encoder-Decoder with RNNs



- Allowing every hidden state of the decoder (not just the first decoder state) to be influenced by by the context c produced by the encoder.
- Context vector c available at each step in the decoding process:

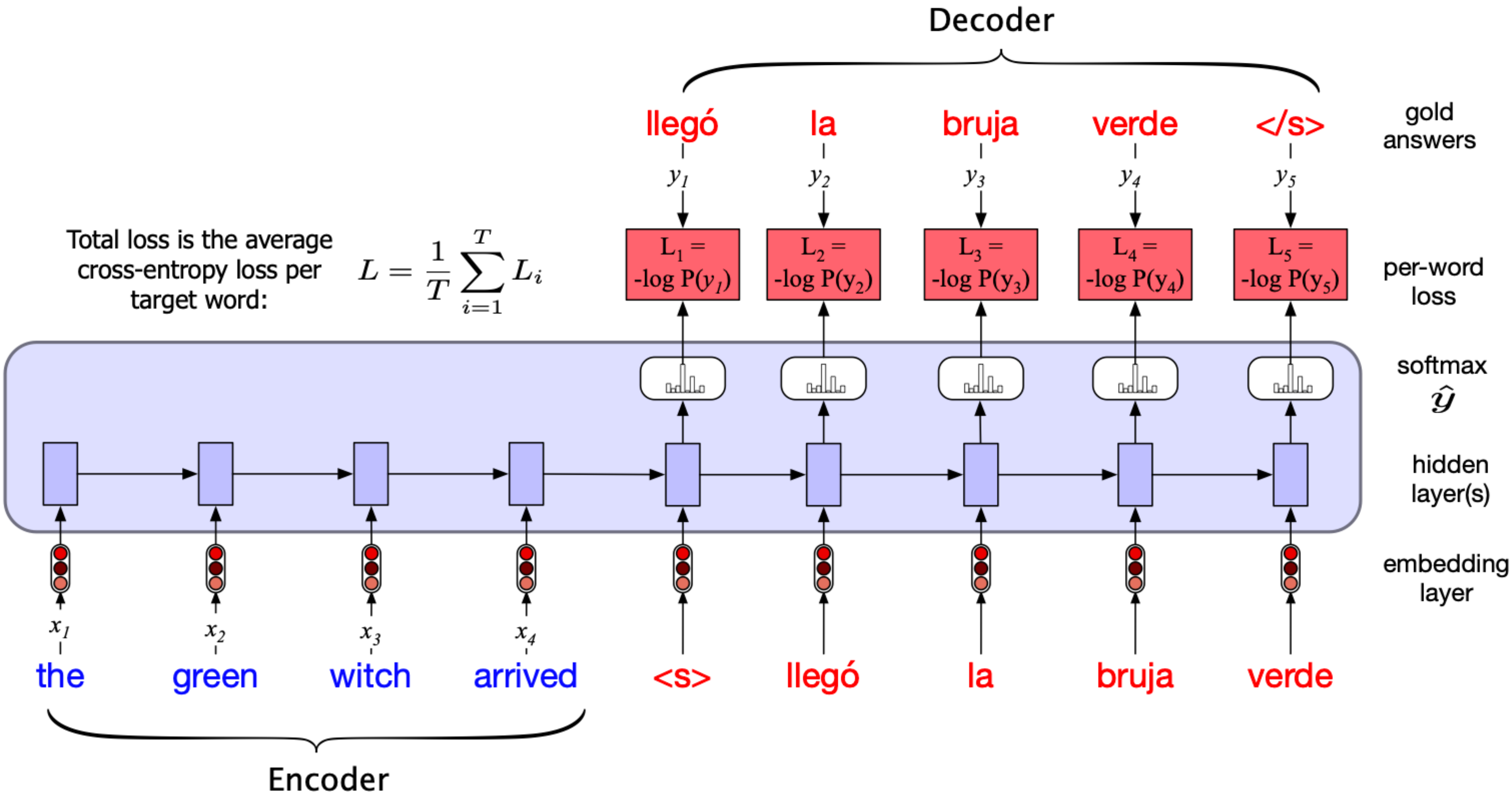
$$\mathbf{h}_t^d = g(\hat{y}_{t-1}, \mathbf{h}_{t-1}^d, \mathbf{c})$$

- Overall:

$$\begin{aligned}
\mathbf{c} &= \mathbf{h}_n^e \\
\mathbf{h}_0^d &= \mathbf{c} \\
\mathbf{h}_t^d &= g(\hat{y}_{t-1}, \mathbf{h}_{t-1}^d, \mathbf{c}) \\
\mathbf{z}_t &= f(\mathbf{h}_t^d) \\
y_t &= \text{softmax}(\mathbf{z}_t)
\end{aligned}$$

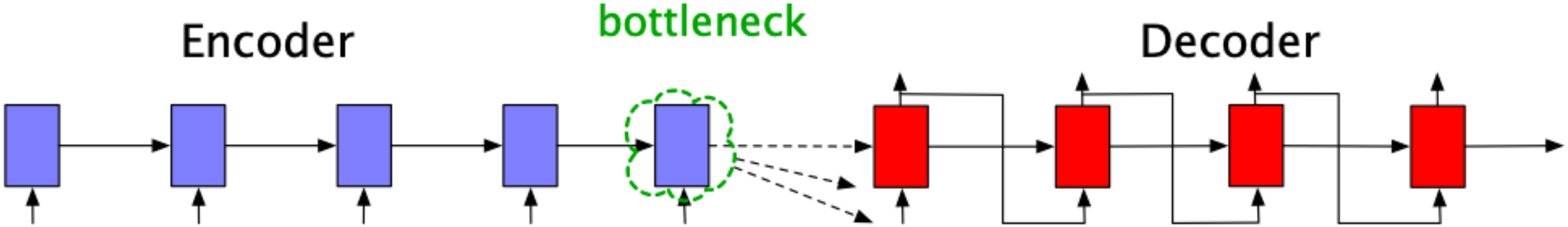
Training Encoder-Decoder with RNNs

- Objective: $\hat{y}_t = \text{argmax}_{w \in V} P(w|x, y_1 \ldots y_{t-1})$



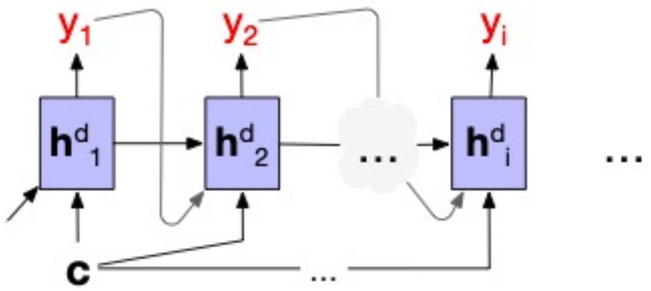
Attention

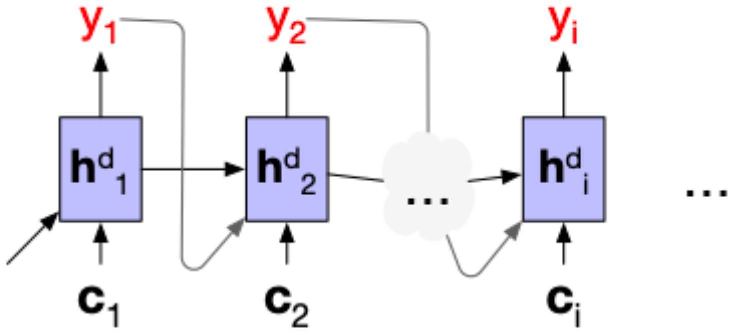
- Requiring the context c to be only the encoder's final hidden state forces all the information from the entire source sentence to pass through this **representational bottleneck**.



Attention

- The attention mechanism allows each hidden state of the decoder to see a different **dynamic context**, which is a **function of all the encoder hidden states.**





$$\mathbf{h}_t^d = g(\hat{y}_{t-1}, \mathbf{h}_{t-1}^d, \mathbf{c})$$

$$\mathbf{h}_i^d = g(\hat{y}_{i-1}, \mathbf{h}_{i-1}^d, \mathbf{c}_i)$$

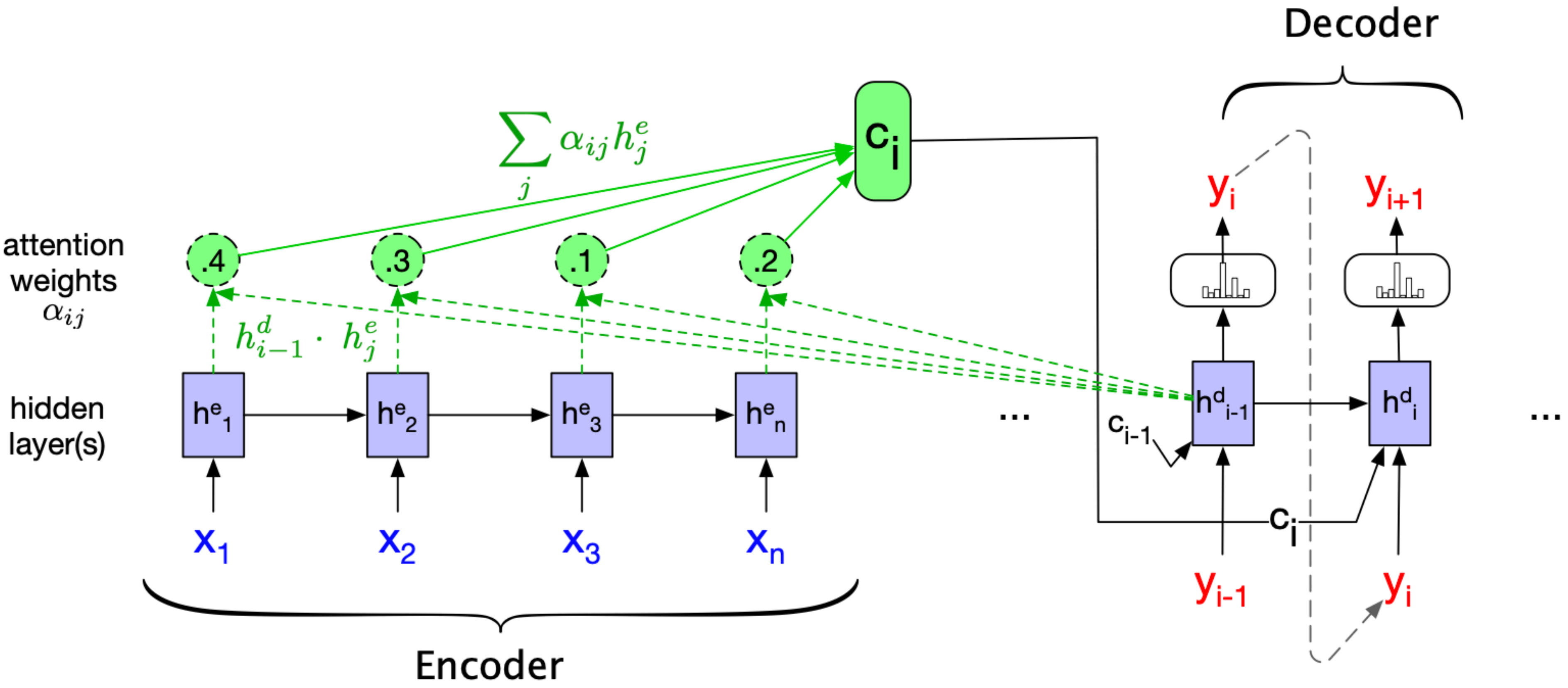
Attention

- Computation of the dynamic context vector:

$$\begin{aligned}\alpha_{ij} &= \mathrm{softmax}(score(\mathbf{h}^d_{i-1}, \mathbf{h}^e_j)\ \forall j \in e) \\ &= \frac{\exp(score(\mathbf{h}^d_{i-1}, \mathbf{h}^e_j)}{\sum_k \exp(score(\mathbf{h}^d_{i-1}, \mathbf{h}^e_k))}\end{aligned}$$

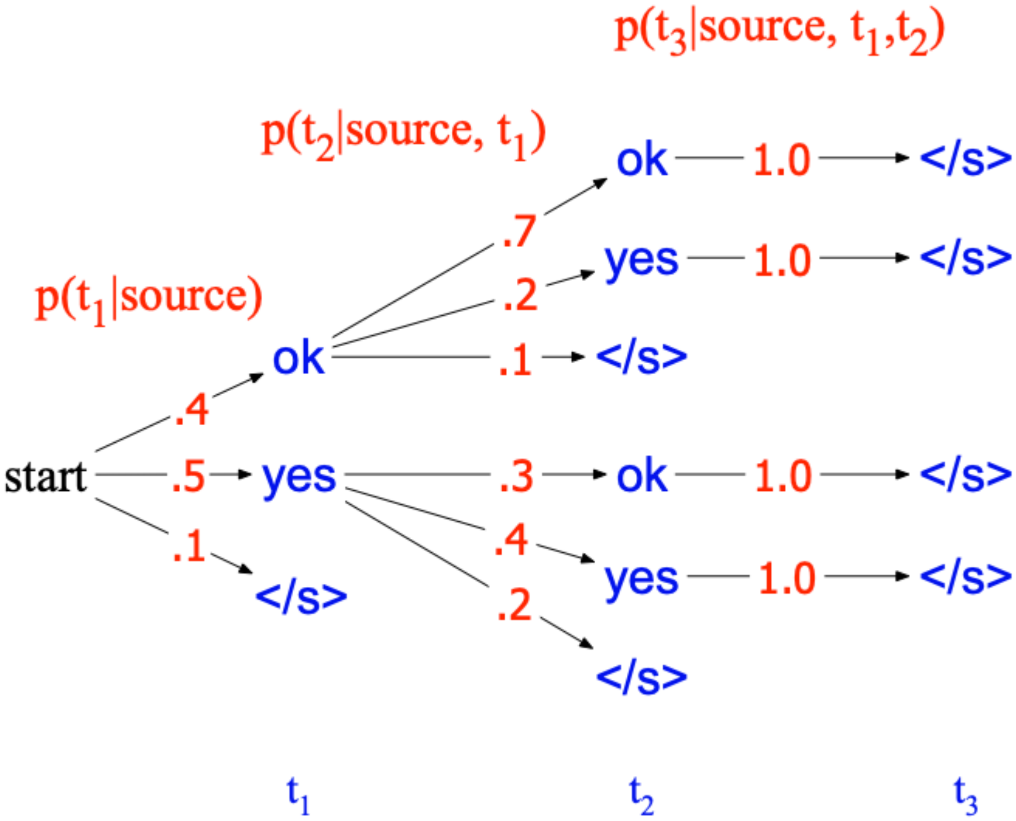
$$\mathbf{c}_i = \sum_j \alpha_{ij}\, \mathbf{h}^e_j$$

Encoder-decoder network with attention

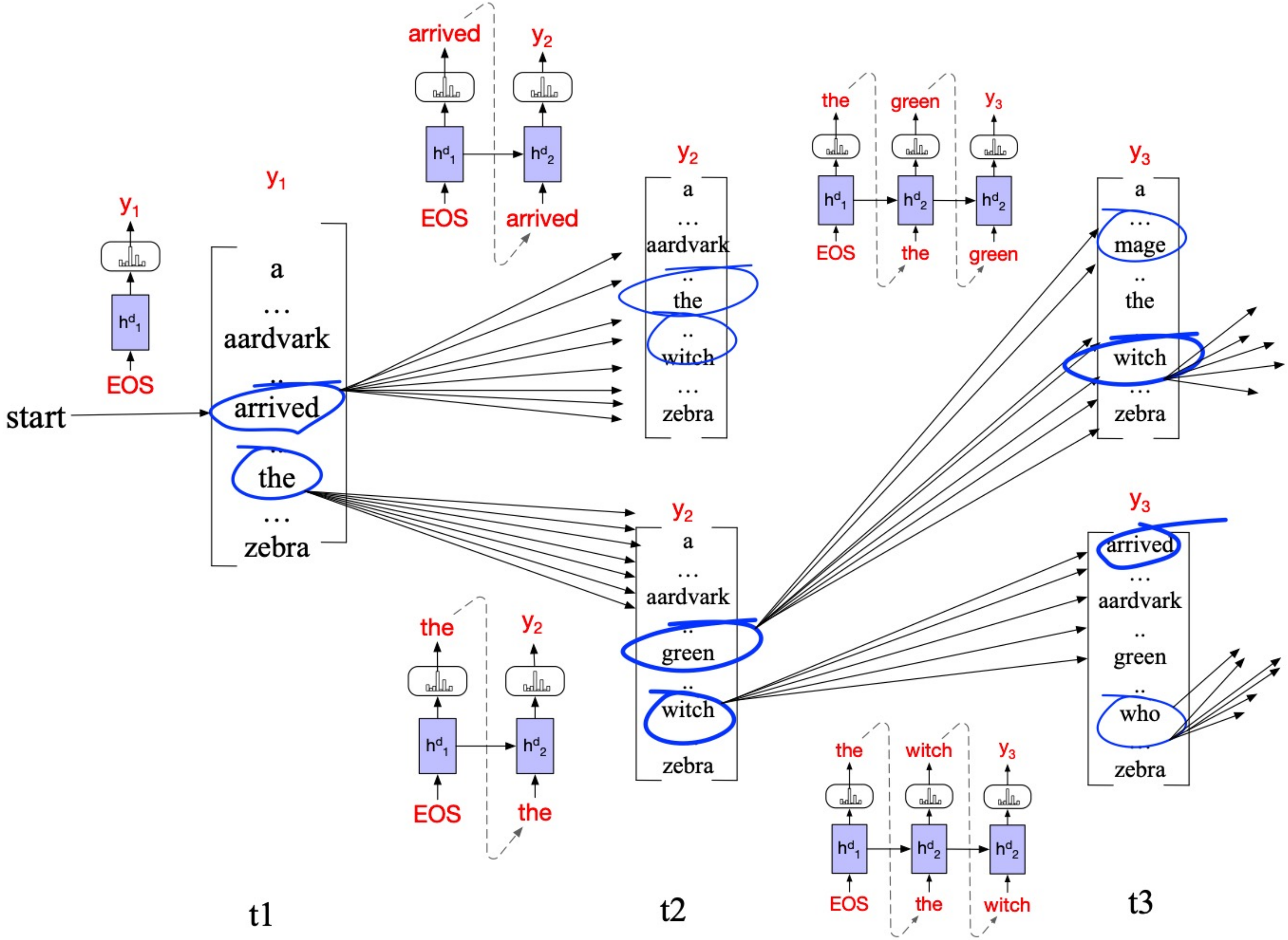


Beam search

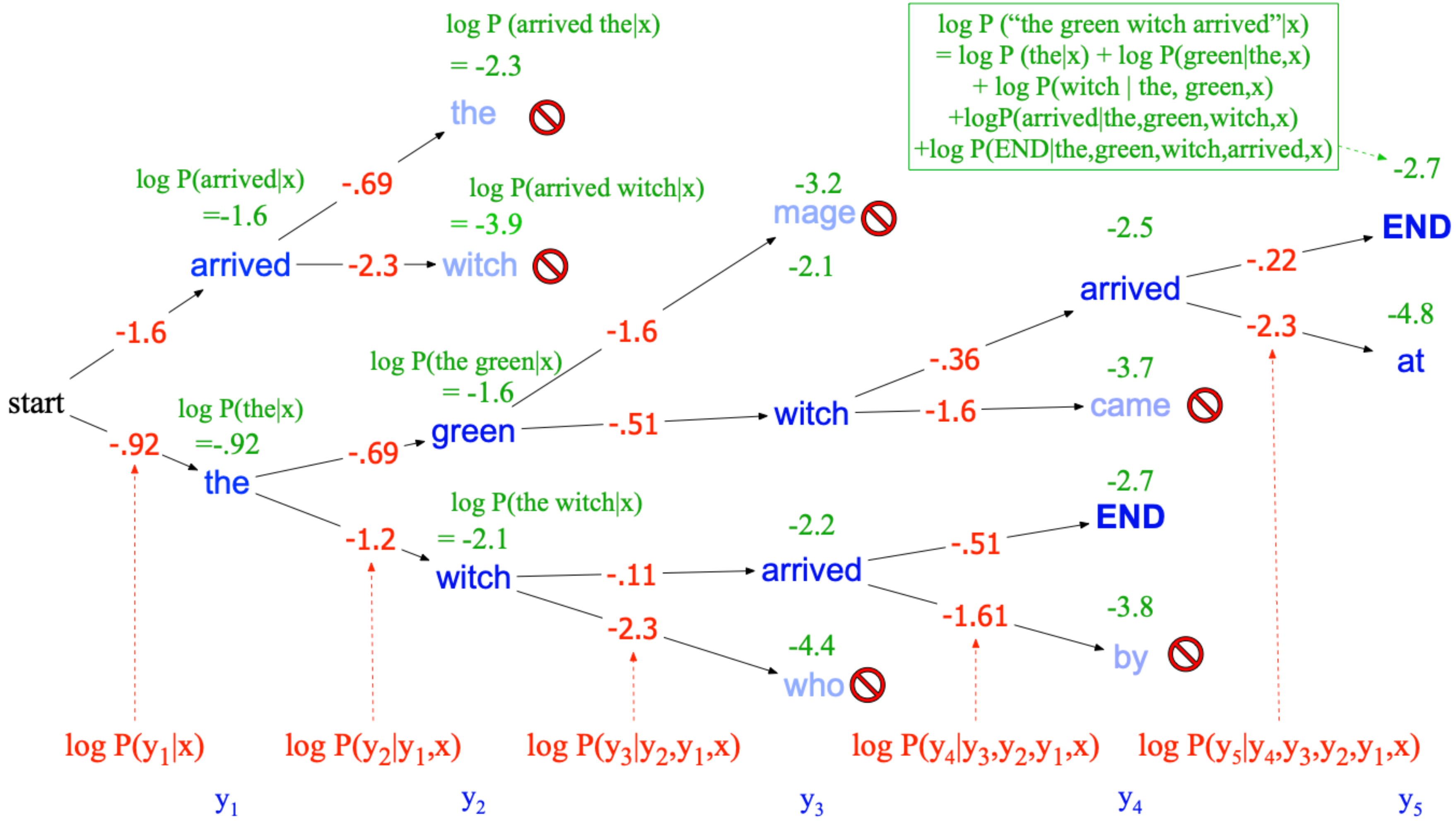
- A search tree for generating the target string T = t_1, t_2, … from the vocabulary V = {yes, ok, <s>}



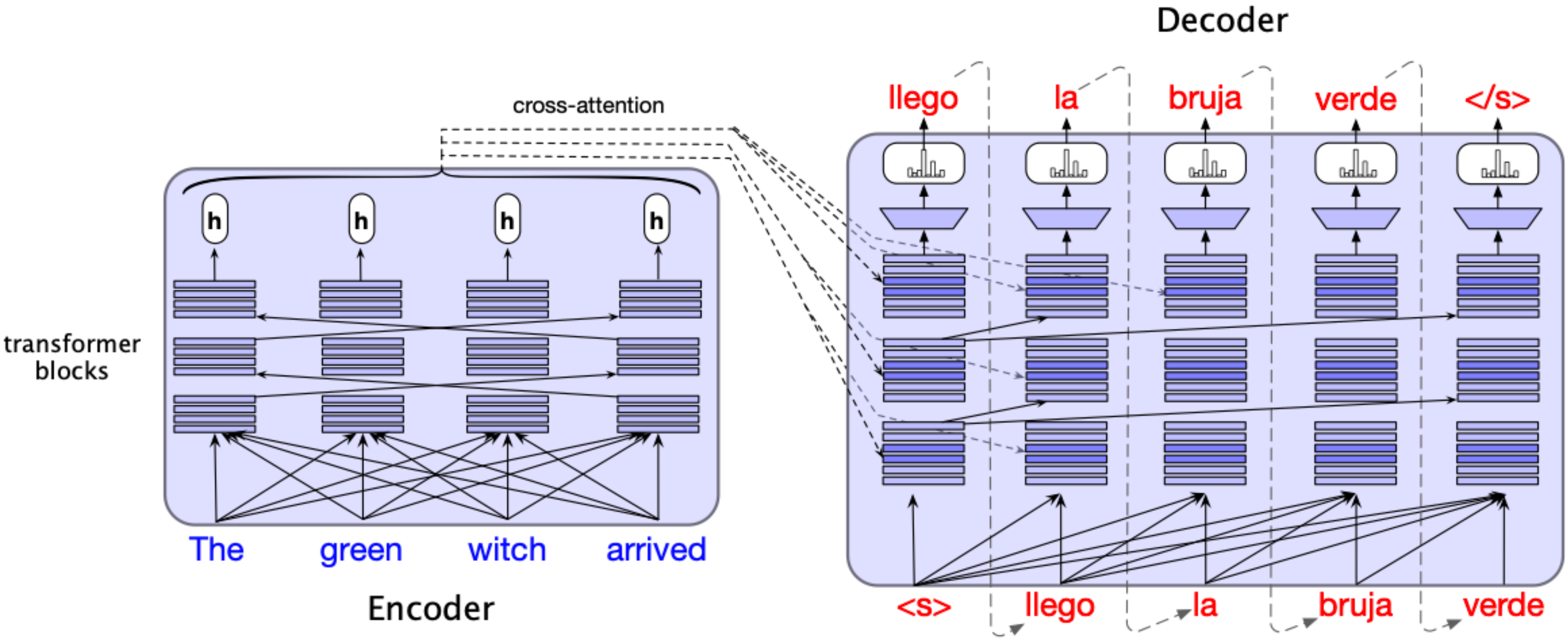
Beam search with a beam width of k = 2



Scoring for beam search with a beam width of k = 2

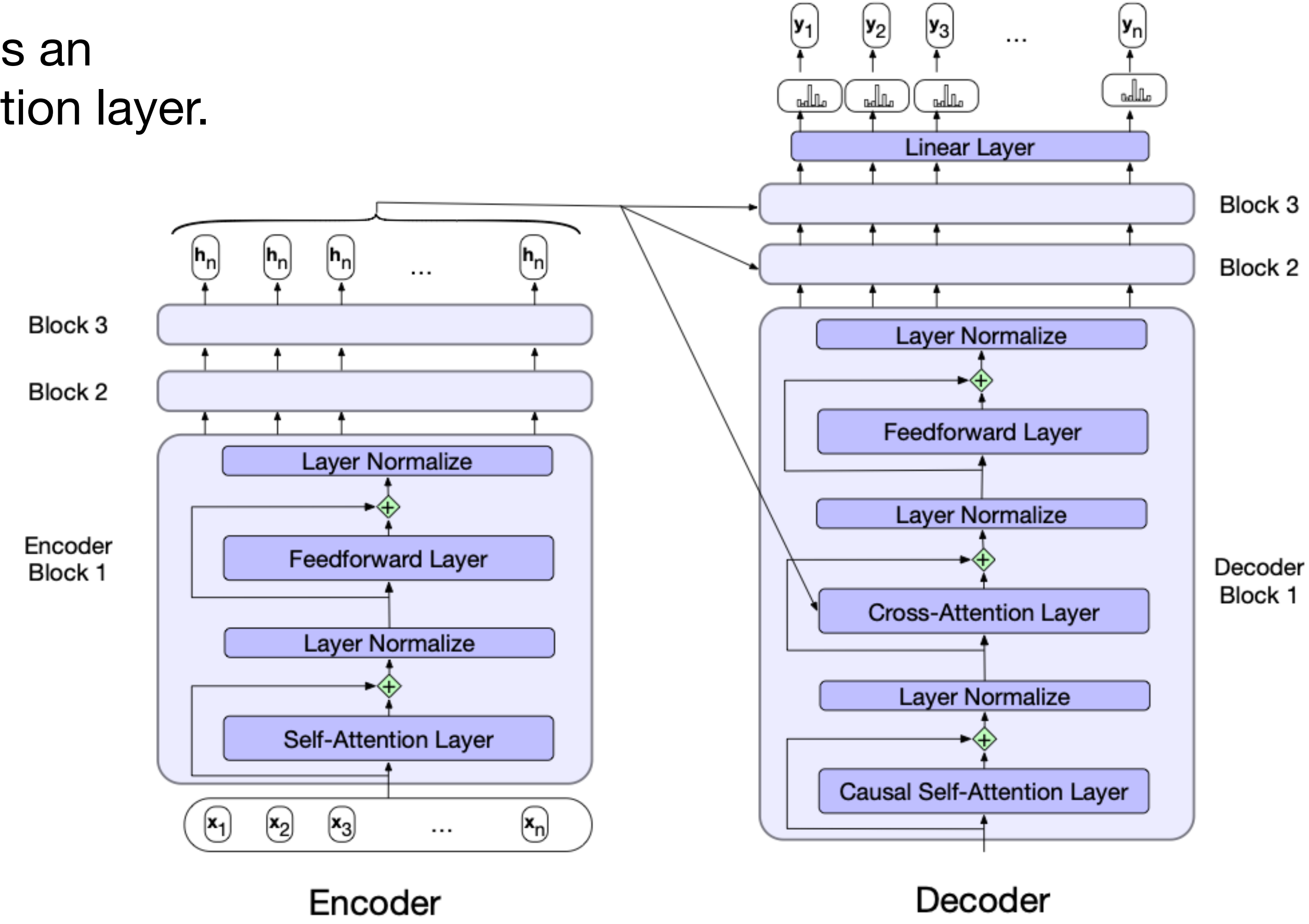


Encoder-Decoder with Transformers



Transformer block for the encoder and decoder

- Each decoder has an extra cross-attention layer.



Rule-based vs. Statistical vs. Neural Machine Translation

Rule-based MT:

- Hand-written transfer rules
- Rules can be based on lexical or structural transfer
- Pro: firm grip on complex translation phenomena
- Con: Often very labor-intensive, lack of robustness

Statistical MT:

- Mainly word or phrase-based translations
- Translation are learned from actual data
- Pro: Translations are learned automatically
- Con: Difficult to model complex translation phenomena

Neural MT: the most recent paradigm (the state-of-the-art as of now).

- End-to-End training: all parameters are simultaneously optimized
- Distributed dense-vector representations: exploits word similarities
- ‘Infinite’ context: neural models better make use of long-range contexts