# COMP 790.139 (Fall 2017)
# Natural Language Processing

## Lecture 7: Summarization; Guest Talk; Machine Translation 1

THE UNIVERSITY
*of* NORTH CAROLINA
*at* CHAPEL HILL

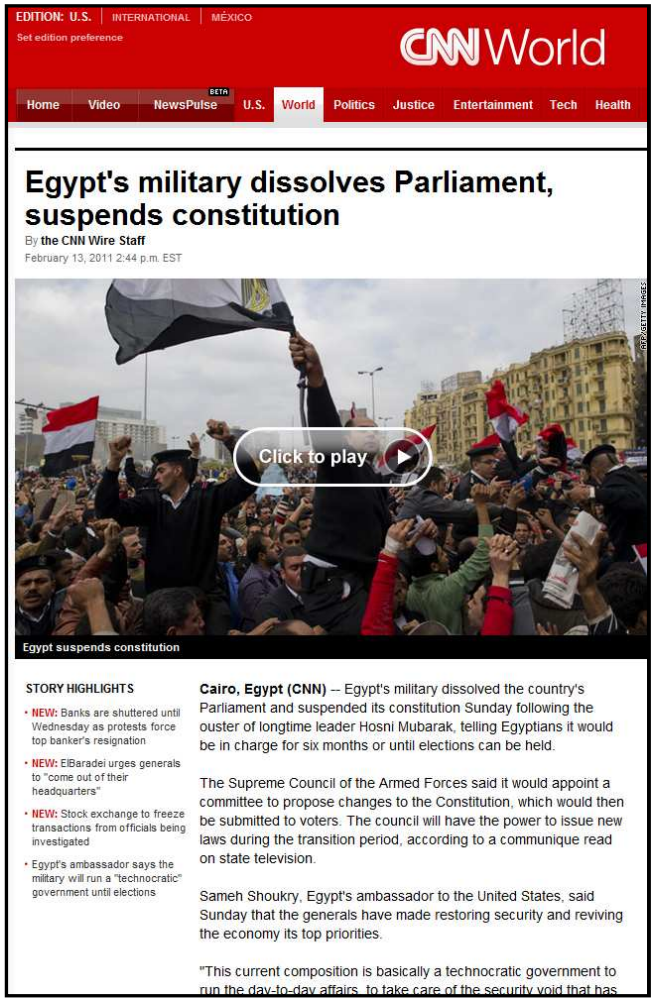
Mohit Bansal

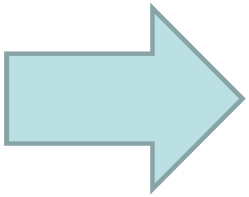
(various slides adapted/borrowed from courses by Dan Klein, JurafskyMartin-SLP3, Manning/Socher, others)

# Automatic Document Summarization

# Single-Document Summarization

- Full document to a salient, non-redundant summary of ~100 words



**STORY HIGHLIGHTS**

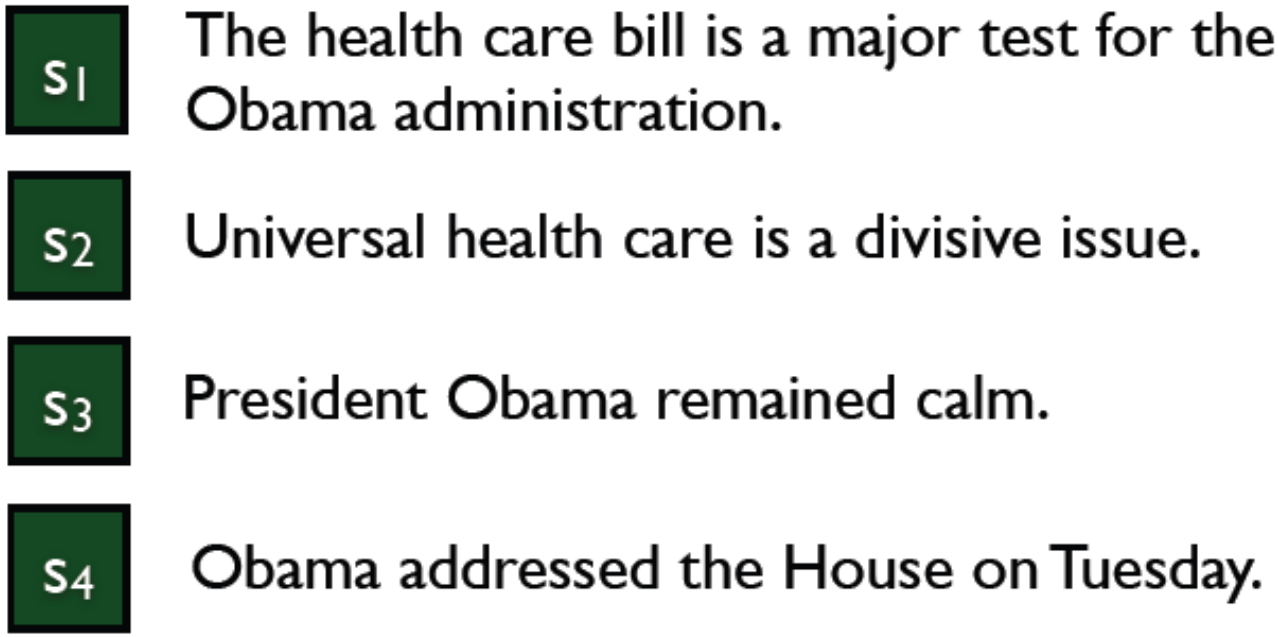
- **NEW:** Banks are shuttered until Wednesday as protests force top banker's resignation
- **NEW:** ElBaradei urges generals to "come out of their headquarters"
- **NEW:** Stock exchange to freeze transactions from officials being investigated
- Egypt's ambassador says the military will run a "technocratic" government until elections

# Multi-Document Summarization

- Several news sources with articles on the same topic (can use overlapping info across articles as a good feature for summarization)

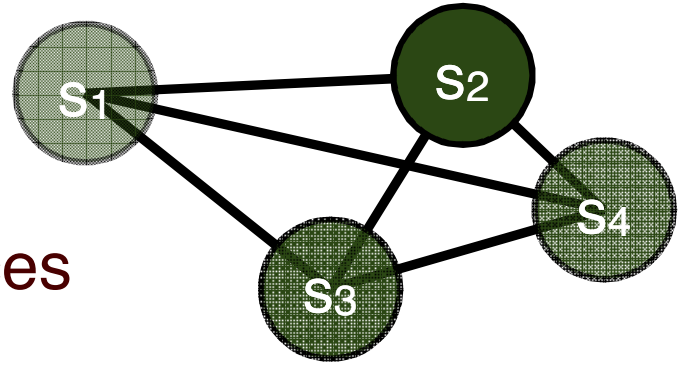
# Extractive Summarization

- Directly selecting existing sentences from input document instead of rewriting them



# Graph-based Extractive Summ

Stationary distribution
represents node centrality




Nodes are sentences

Edges are similarities

[Mihalcea et al., 2004, 2005; inter alia]

# Maximize Concept Coverage

$s_1$ The health care bill is a major test for the Obama administration.

$s_2$ Universal health care is a divisive issue.

$s_3$ President Obama remained calm.

$s_4$ Obama addressed the House on Tuesday.

| concept | value |
|---|---|
| obama | 3 |
| health | 2 |
| house | 1 |

Length limit:
18 words

| summary | length | value | |
|---|---|---|---|
| $\{s_1, s_3\}$ | 17 | 5 | ← greedy |
| $\{s_2, s_3, s_4\}$ | 17 | 6 | ← optimal |

[Gillick and Favre, 2009]

# Maximize Concept Coverage

- A set coverage optimization problem

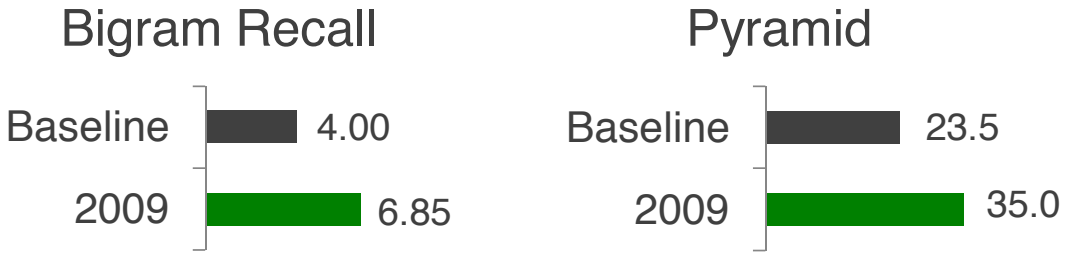
$$\max_{s \in S(D)} \sum_{c \in C(s)} v_c$$

Value of concept c

Set of extractive summaries of document set D

Set of concepts present in summary s

## Results




[Gillick and Favre, 2009]

# Maximize Concept Coverage

- Can be solved using an integer linear program with constraints:

$$\text{Maximize: } \sum_i w_i c_i$$ ← total concept value

$$\text{Subject to: } \sum_j l_j s_j \leq L$$ ← summary length limit

$$s_j Occ_{ij} \leq c_i, \quad \forall i, j$$

$$\sum_j s_j Occ_{ij} \geq c_i \quad \forall i$$

← maintain consistency between selected sentences and concepts

$$c_i \in \{0, 1\} \quad \forall i$$

$$s_j \in \{0, 1\} \quad \forall j$$

[Gillick et al., 2008] [Gillick and Favre, 2009]

# Beyond Extraction: Compression

- If you had to write a concise summary, making effective use of the 100-word limit, you would remove some information from the lengthy sentences in the original article

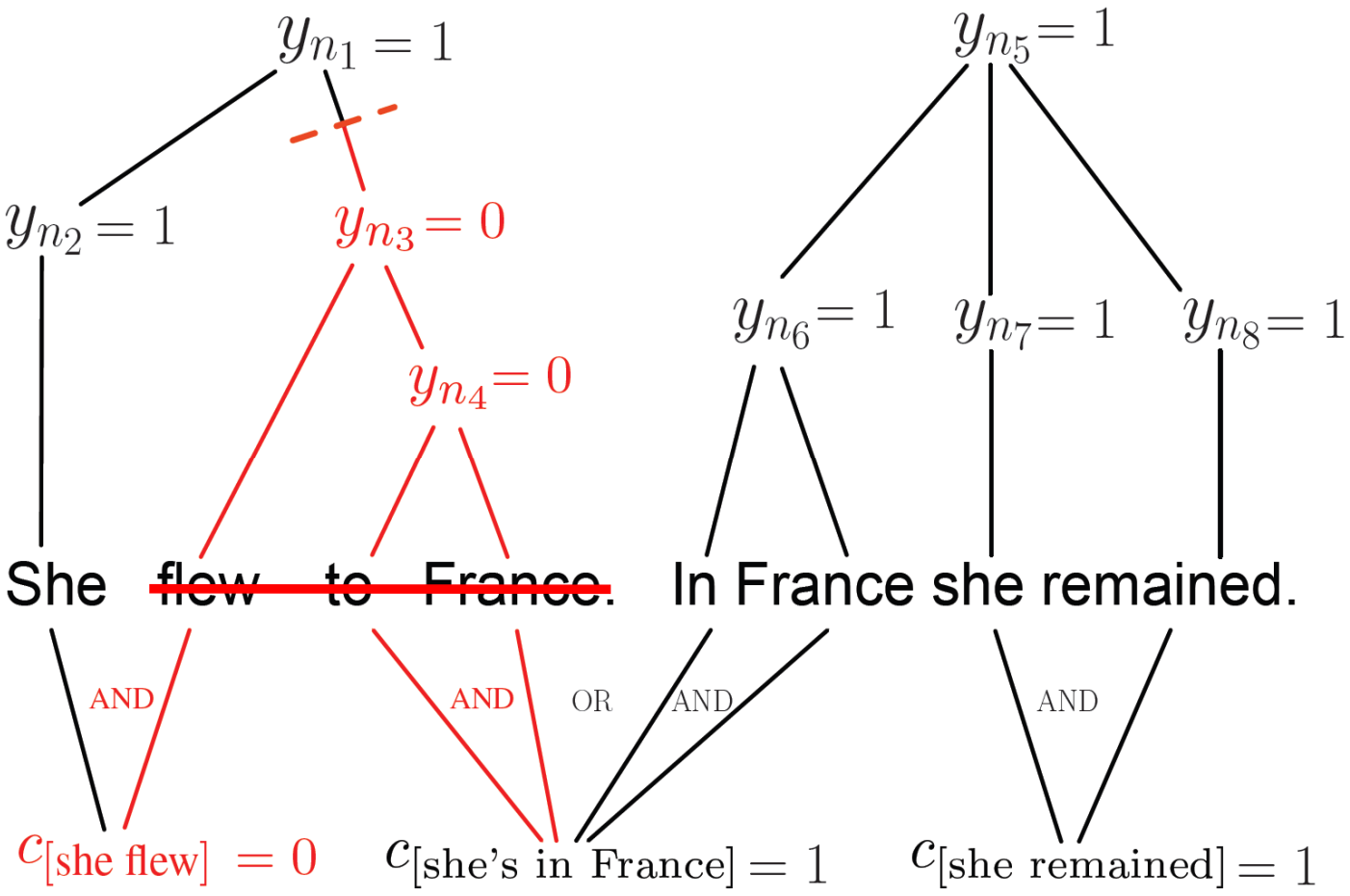
What would a human do?

~~It is therefore unsurprising that~~ Lindsay pleaded not guilty ~~yesterday afternoon~~ to the charges filed against her, ~~according to her publicist.~~

[Berg-Kirkpatrick et al., 2011]

# Beyond Extraction: Compression

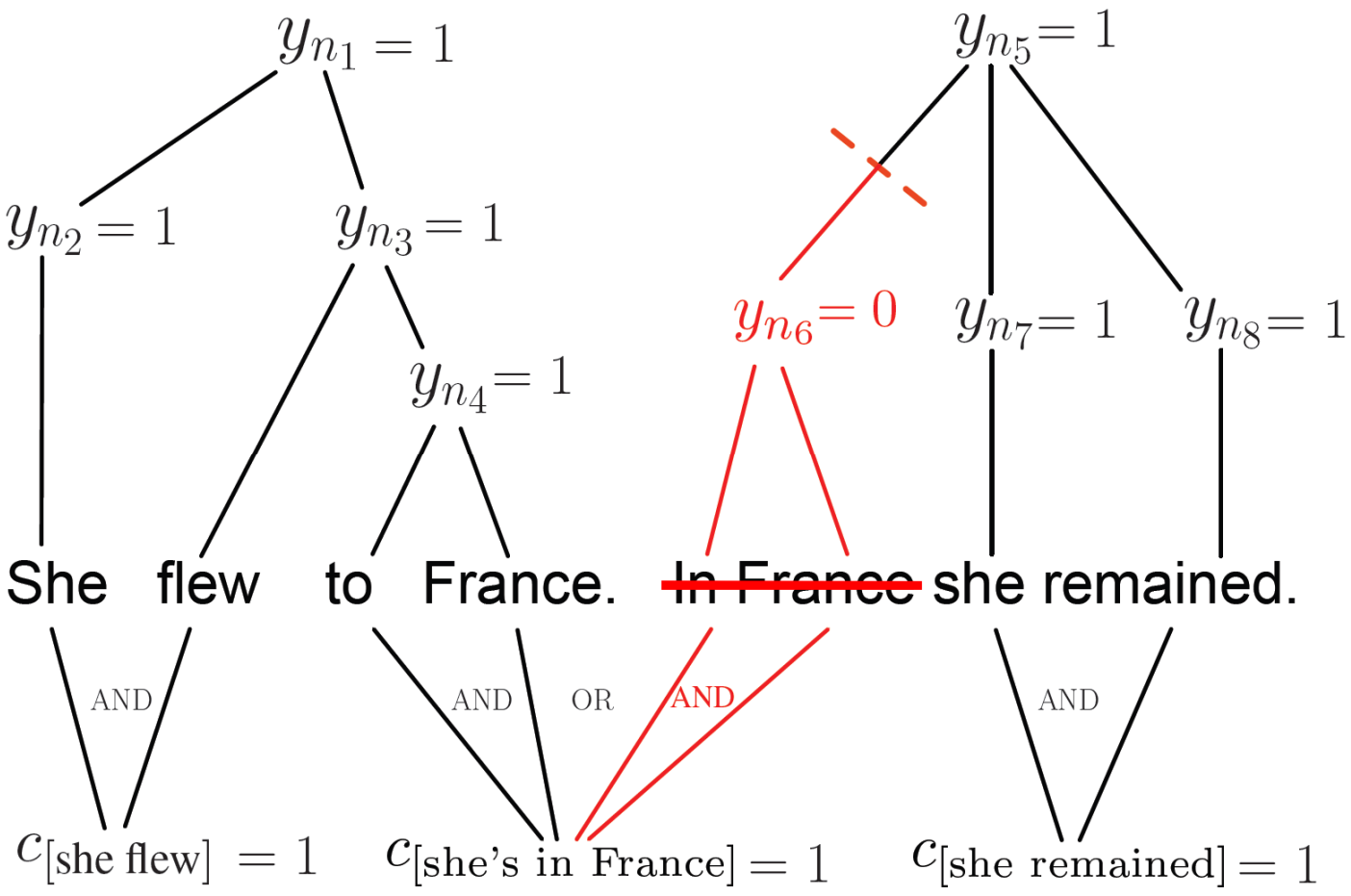
- Model should learn the subtree deletions/cuts that allow compression




[Berg-Kirkpatrick et al., 2011]

# Beyond Extraction: Compression

- Model should learn the subtree deletions/cuts that allow compression




[Berg-Kirkpatrick et al., 2011]

# Beyond Extraction: Compression

- The new optimization problem looks to maximize the concept values as well as safe deletion values in the candidate summary:

$$\max_{s \in S(D)} \left[ \sum_{c \in C(s)} v_c + \sum_{d \in D(s)} v_d \right]$$

Value of deletion d

Set branch cut deletions made in creating summary s

- To decide the value/cost of a deletion, we decide relevant deletion features and the model learns their weights:

$$v_d = w^\top f(d)$$

[Berg-Kirkpatrick et al., 2011]

# Beyond Extraction: Compression

- Some example features for concept bigrams and cuts/deletions:

## Bigram Features $f(b)$

| | |
|---|---|
| COUNT: | Bucketed document counts |
| STOP: | Stop word indicators |
| POSITION: | First document position indicators |
| CONJ: | All two- and three-way conjunctions of above |
| BIAS: | Always one |

## Cut Features $f(c)$

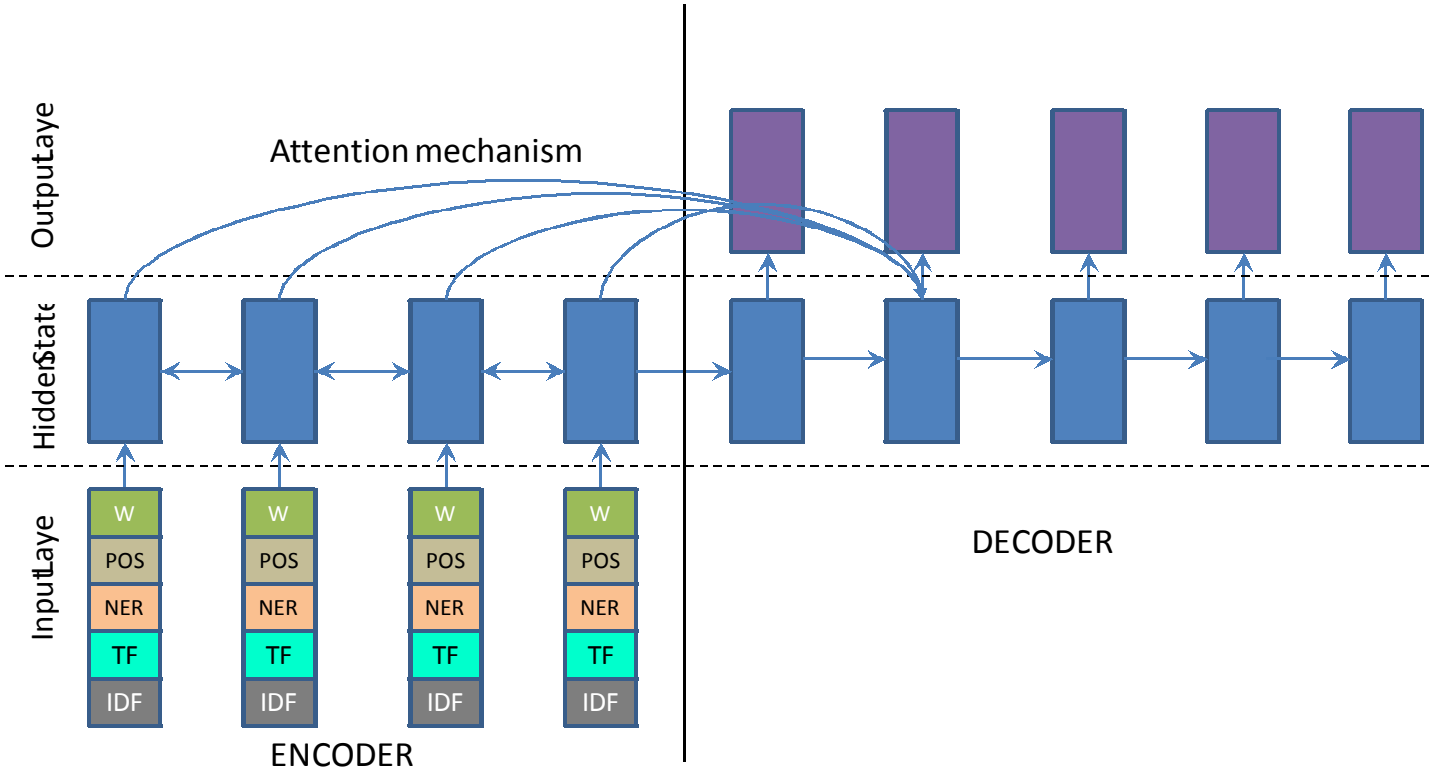
| | |
|---|---|
| COORD: | Coordinated phrase, four versions: NP, VP, S, SBAR |
| S-ADJUNCT: | Adjunct to matrix verb, four versions: CC, PP, ADVP, SBAR |
| REL-C: | Relative clause indicator |
| ATTR-C: | Attribution clause indicator |
| ATTR-PP: | PP attribution indicator |
| TEMP-PP: | Temporal PP indicator |
| TEMP-NP | Temporal NP indicator |
| BIAS: | Always one |

[Berg-Kirkpatrick et al., 2011]

# Neural Abstractive Summarization

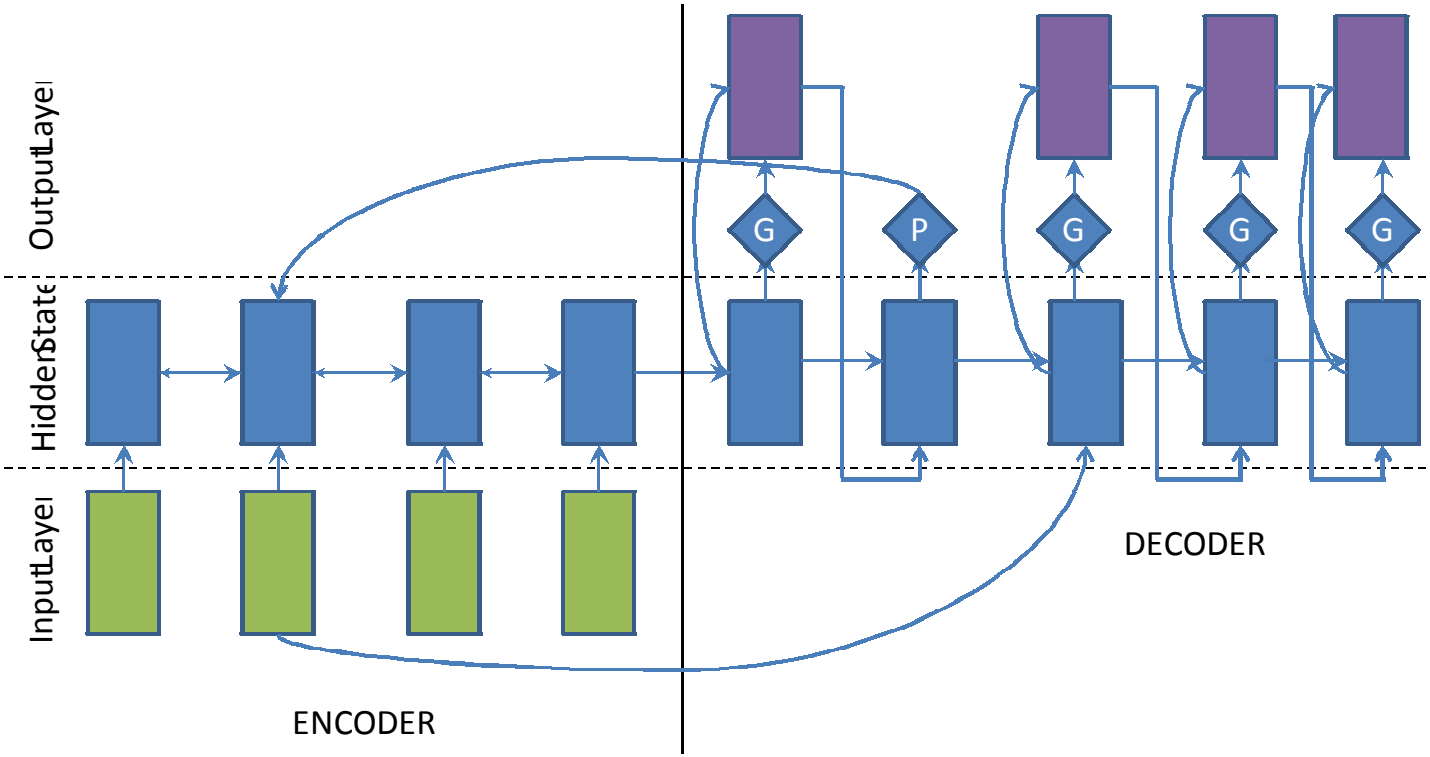
- Mostly based on sequence-to-sequence RNN models
- Later added attention, coverage, pointer/copy, hierarchical encoder/attention, metric rewards RL, etc.
- Examples: Rush et al., 2015; Nallapati et al., 2016; See et al., 2017; Paulus et al., 2017

# Feature-Augmented Encoder-Decoder



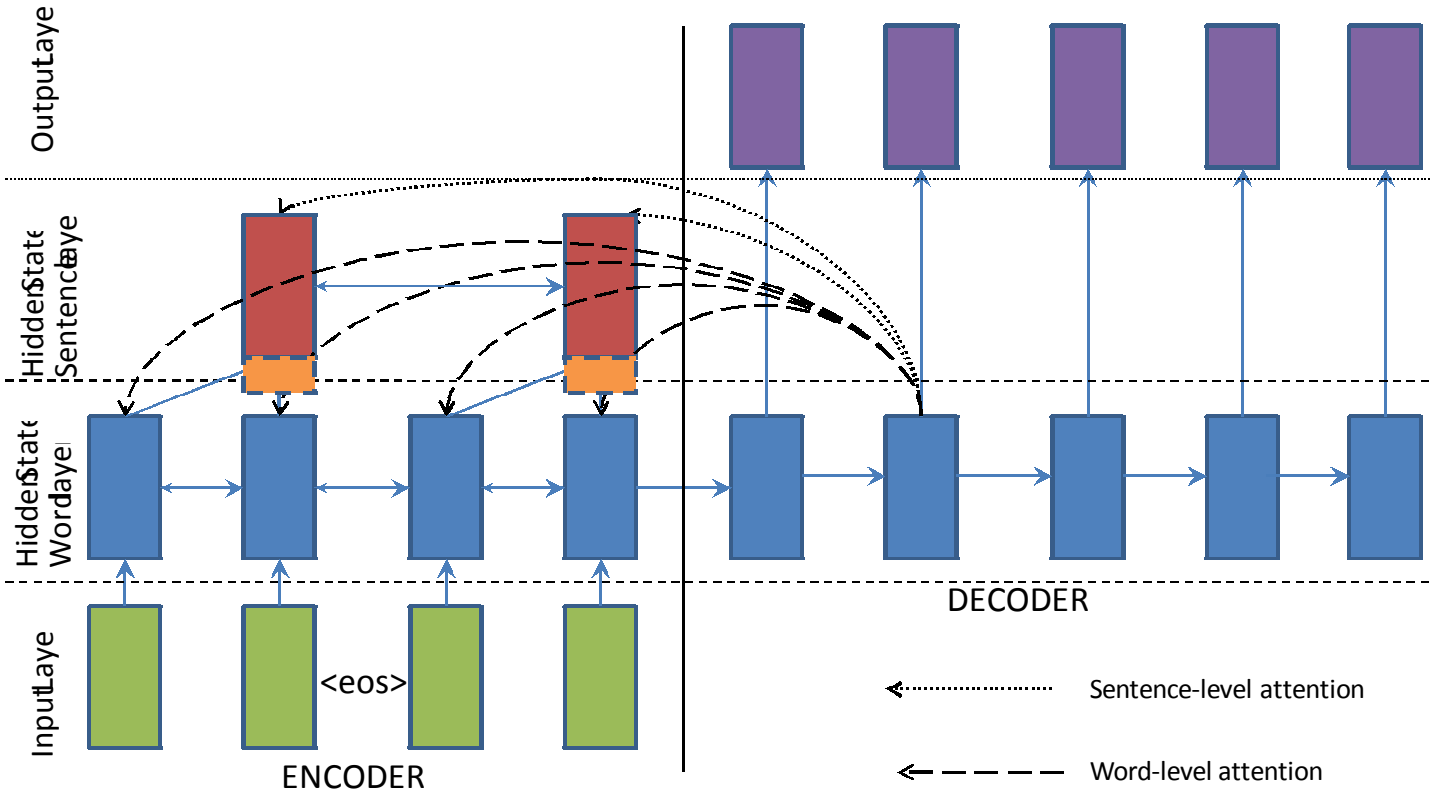

[Nallapati et al., 2016]

# Generation+Copying



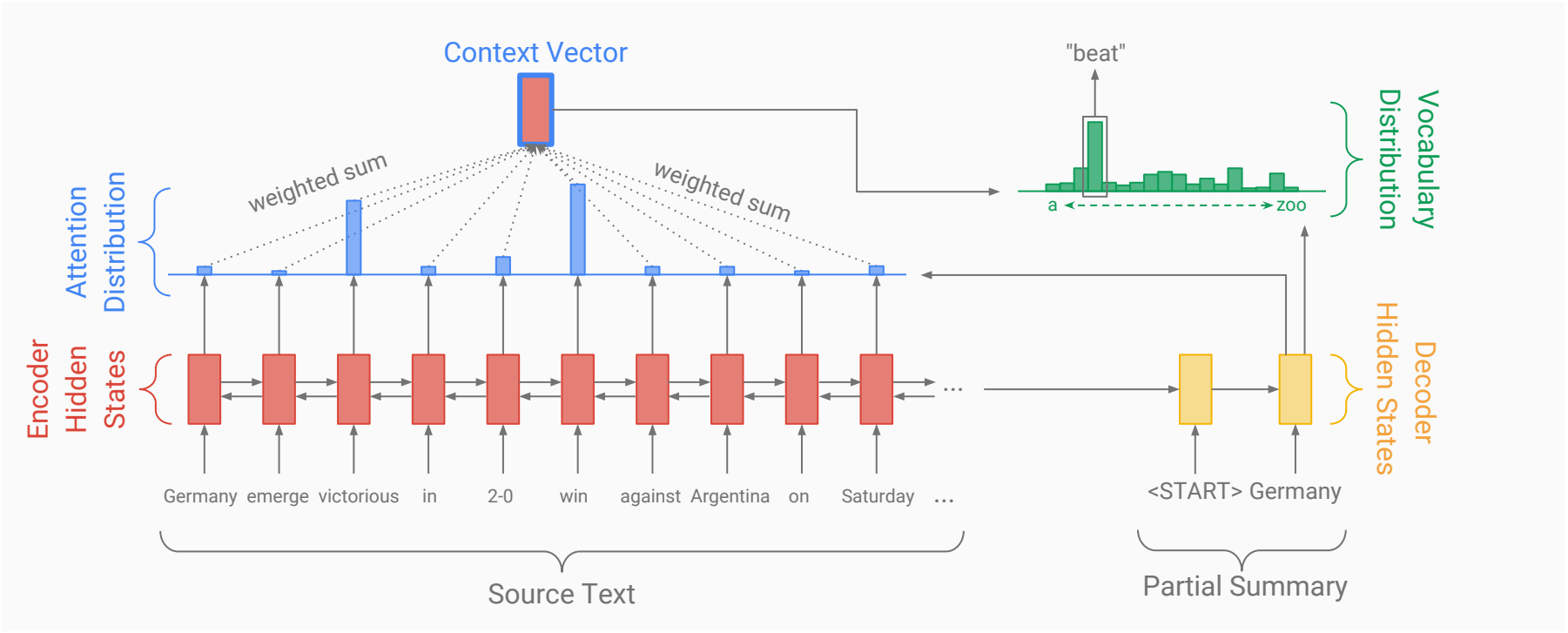

[Nallapati et al., 2016]

# Hierarchical Attention



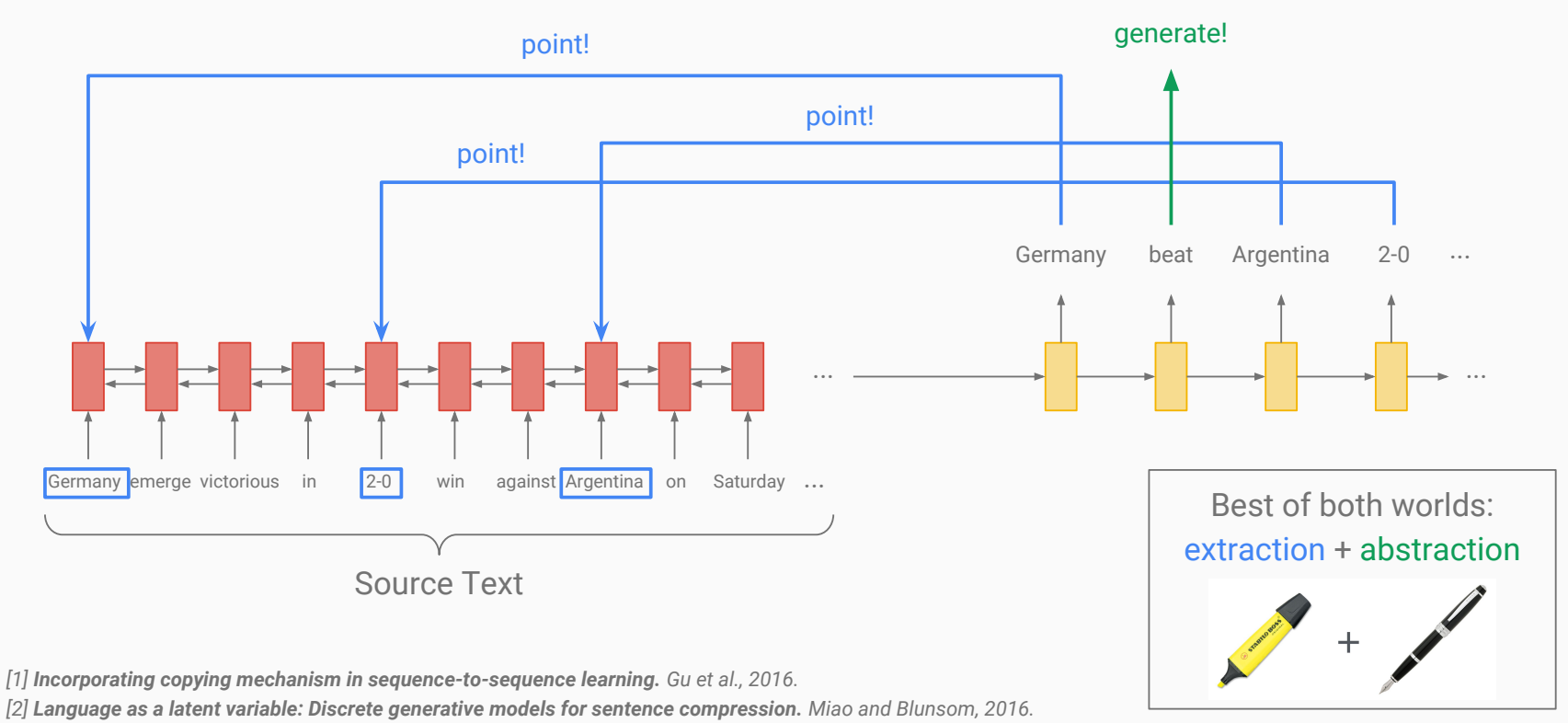

[Nallapati et al., 2016]

# Pointer-Generator Networks




[See et al., 2017]

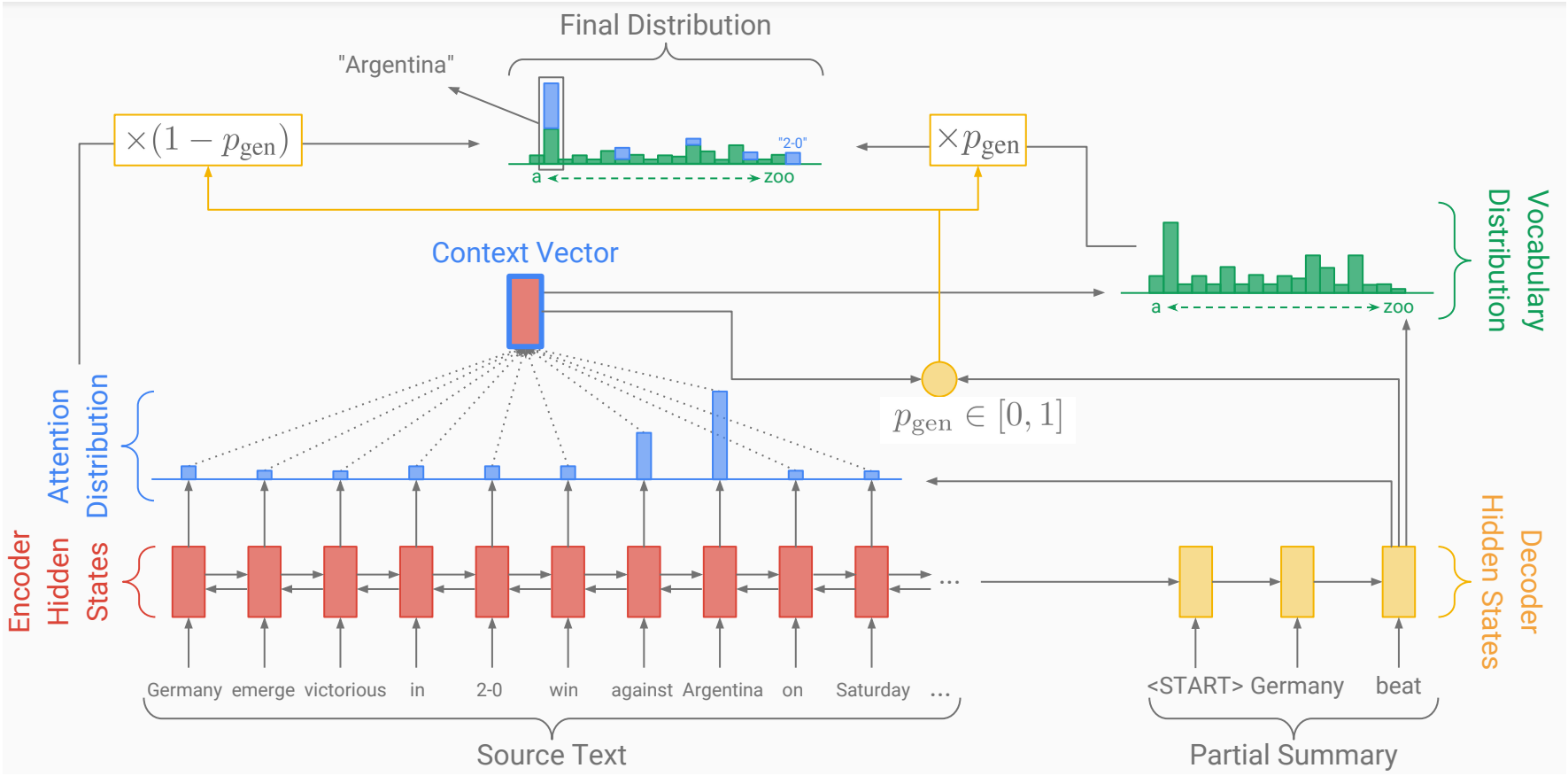
# Pointer-Generator Networks




[1] ***Incorporating copying mechanism in sequence-to-sequence learning.*** *Gu et al., 2016.*
[2] ***Language as a latent variable: Discrete generative models for sentence compression.*** *Miao and Blunsom, 2016.*

[See et al., 2017]

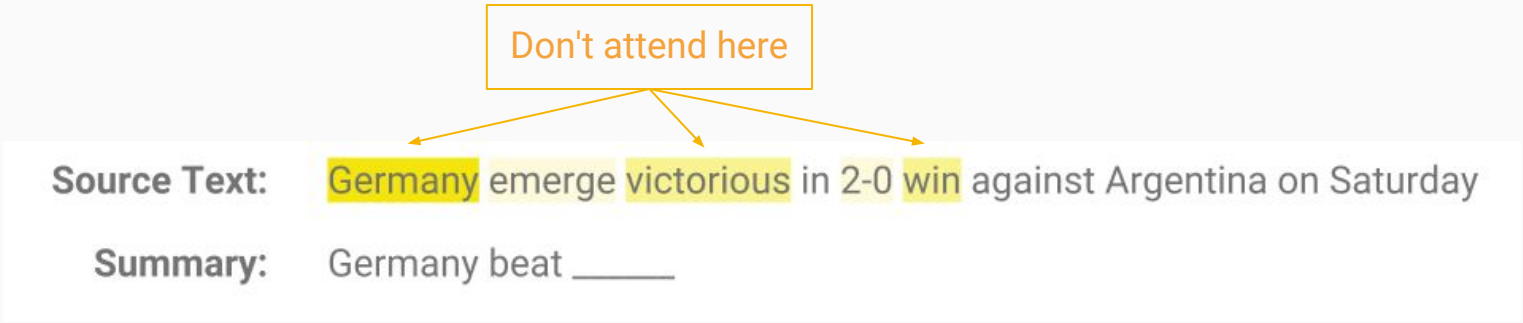
# Pointer-Generator Networks




[See et al., 2017]

# Coverage for Redundancy Reduction

*Coverage* = cumulative attention = what has been covered so far




1. Use coverage as extra input to attention mechanism.
2. Penalize attending to things that have already been covered.

**Result:** repetition rate reduced to level similar to human summaries


*[4]* ***Modeling coverage for neural machine translation.*** *Tu et al., 2016,*
*[5]* ***Coverage embedding models for neural machine translation.*** *Mi et al., 2016*
*[6]* ***Distraction-based neural networks for modeling documents.*** *Chen et al., 2016.*


[See et al., 2017]

# Guest Talk by Ramakanth Pasunuru:

## “Towards Improving Abstractive Summarization via Entailment Generation”

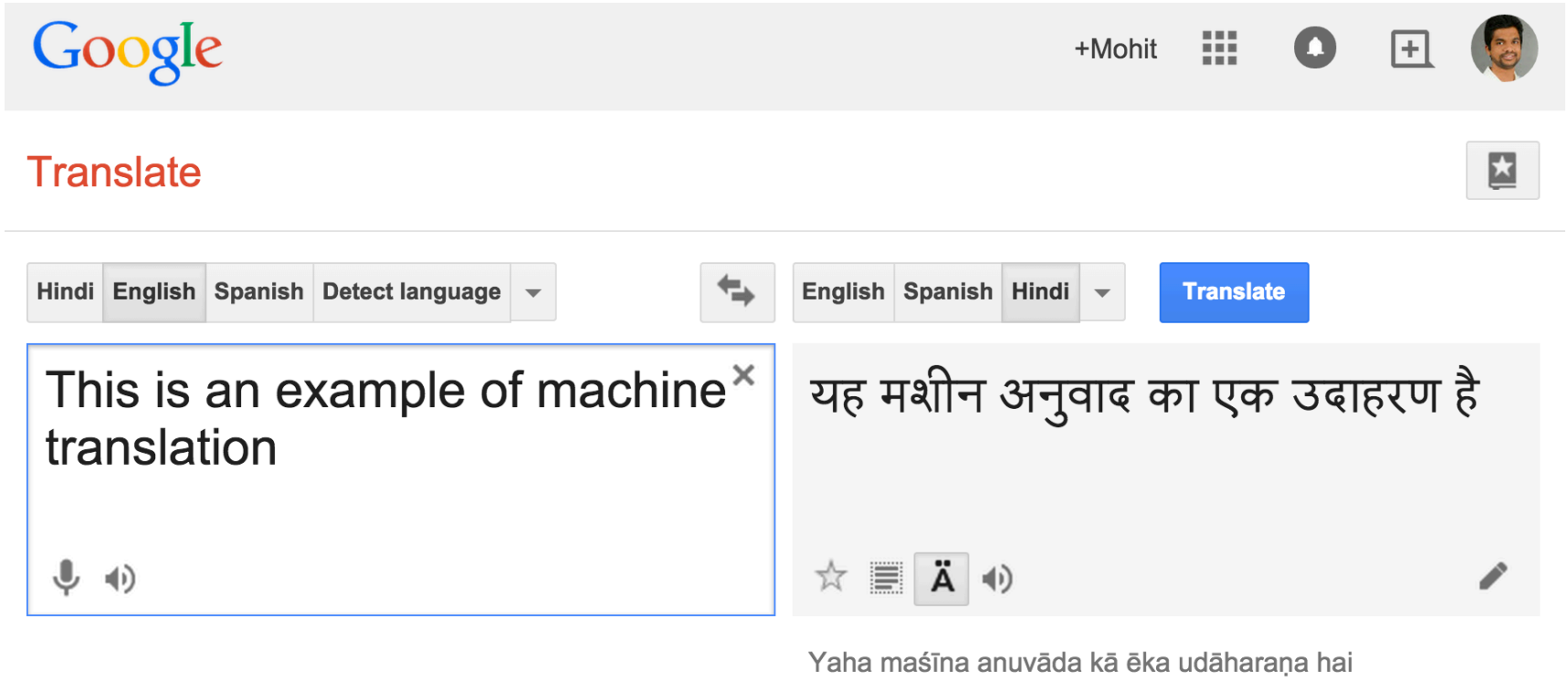
(30 mins)

[Pasunuru, Guo, Bansal. New Summarization Frontiers Workshop, EMNLP 2017]

# Machine Translation

# Machine Translation

- Useful for tons of companies, online traffic, and our international communication!



# Statistical Machine Translation

- Source language f (e.g., French)
- Target language e (e.g., English)
- We want the best target (English) translation given the source (French) input sentence, hence the probabilistic formulation is:

$$\hat{e} = \text{argmax}_e p(e|f)$$

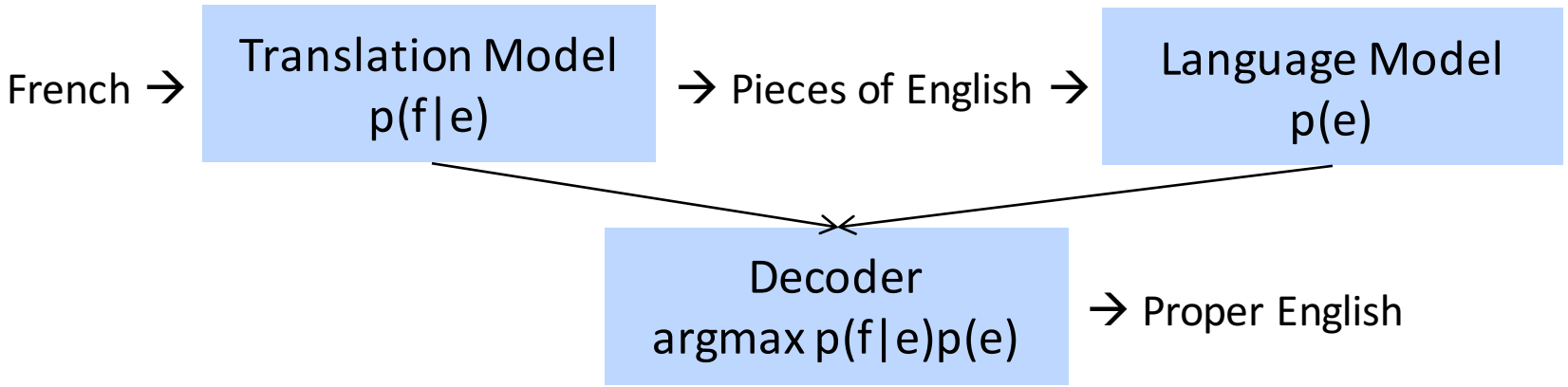
- Using Bayes rule, we get the following (since p(f) in the denominator is independent of the argmax over e):

$$\hat{e} = \text{argmax}_e p(e|f) = \text{argmax}_e p(f|e)p(e)$$

[Richard Socher CS224d]

# Statistical Machine Translation

- The first part is known as the 'Translation Model' p(f|e) and is trained on parallel corpora of {f,e} sentence pairs, e.g., from EuroParl or Canadian parliament proceedings in multiple languages
- The second part p(e) is the 'Language Model' and can be trained on tons more monolingual data, which is much easier to find!

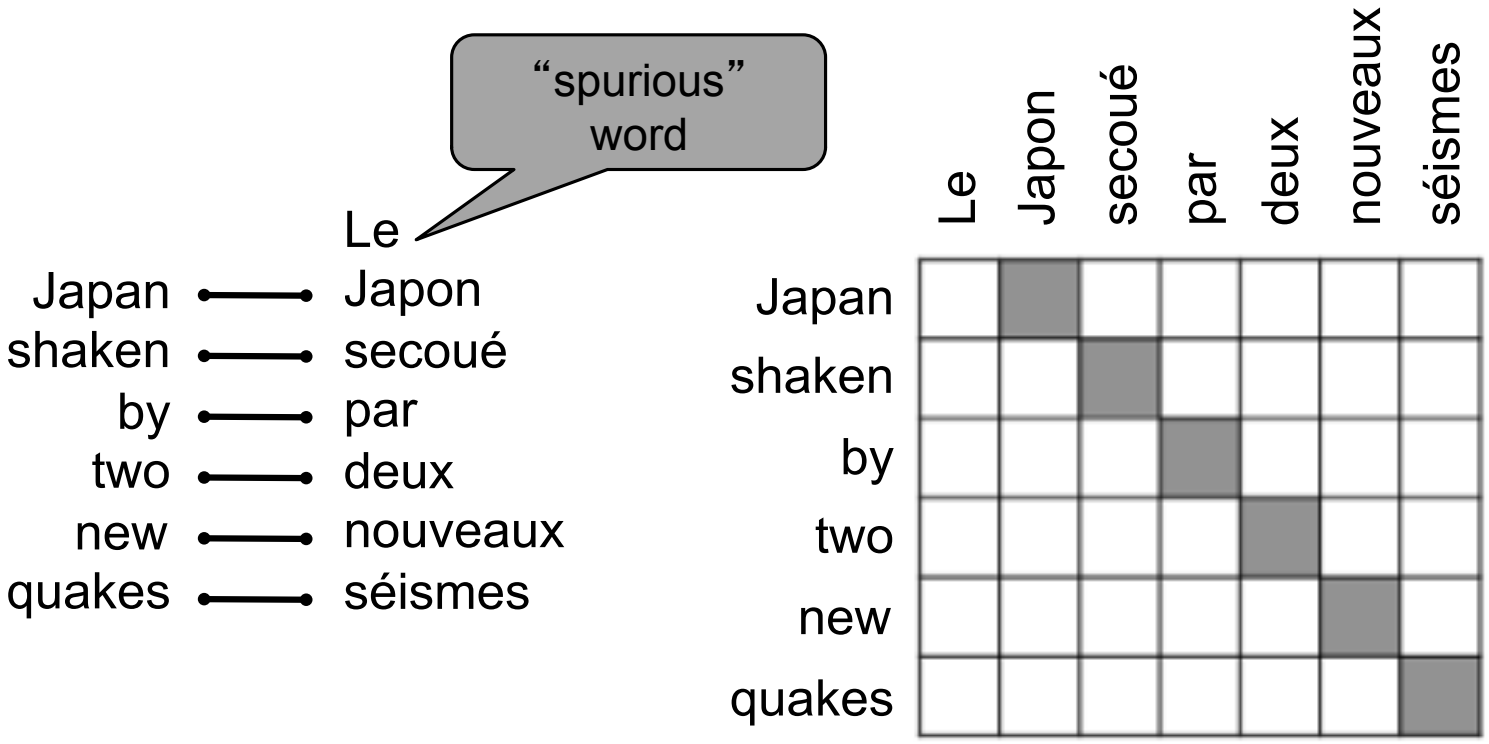



[Richard Socher CS224d]

# Statistical Machine Translation

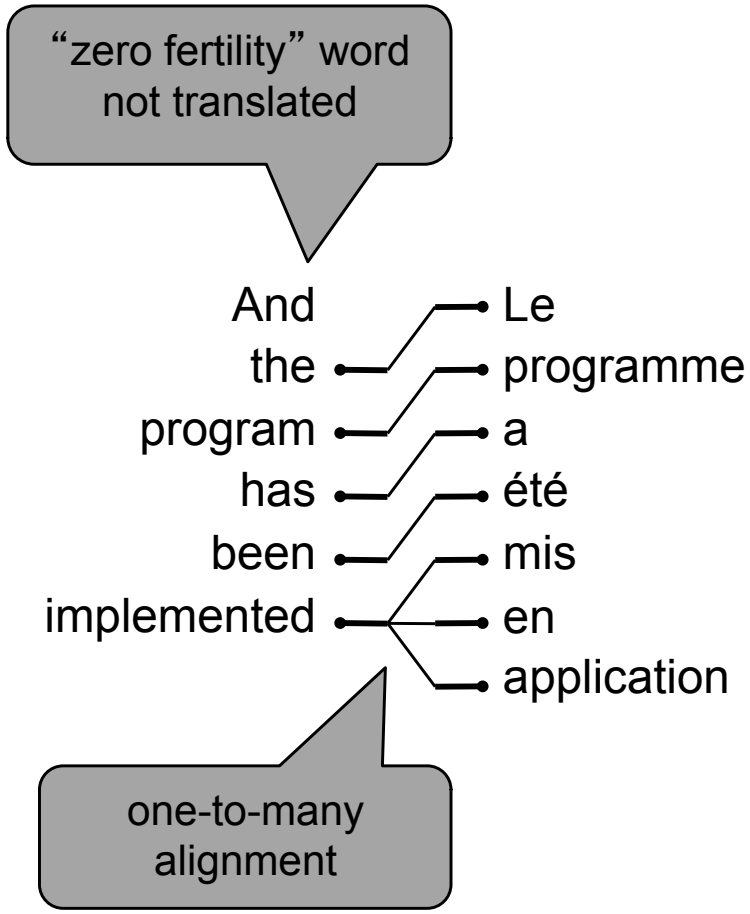
- First step in traditional machine translation is to find alignments or translational matchings between the two sentences, i.e., predict which words/phrases in French align to which words/phrases in English.

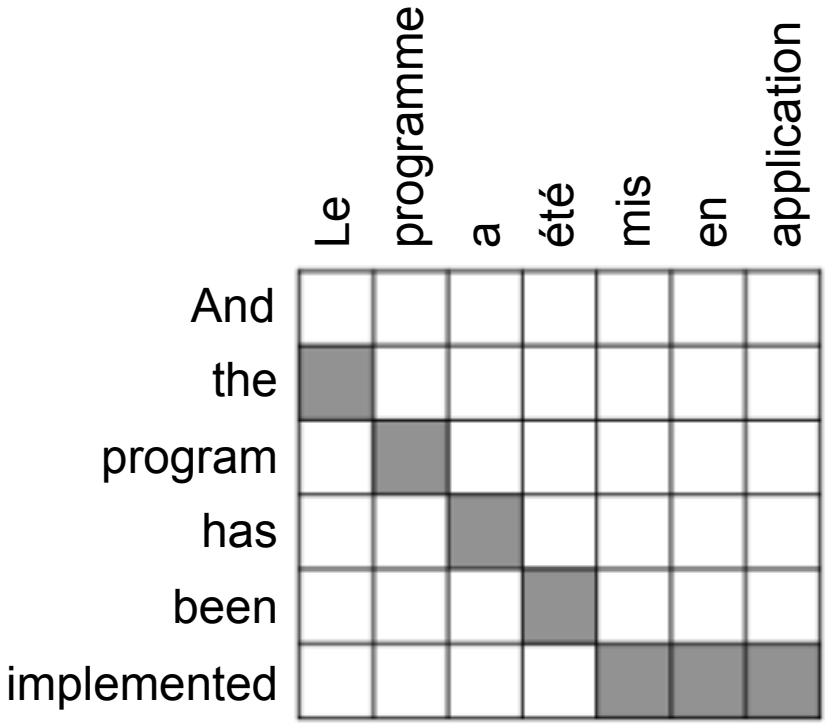
- Challenging problem: e.g., some words may not have any alignments:



# Statistical Machine Translation

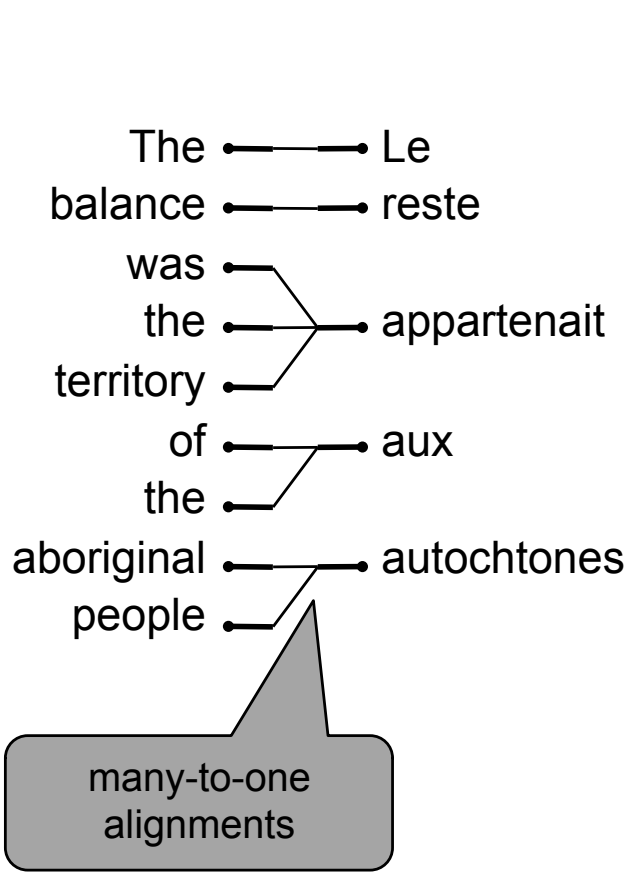
- One word in the source sentence might align to several words in the target sentence:

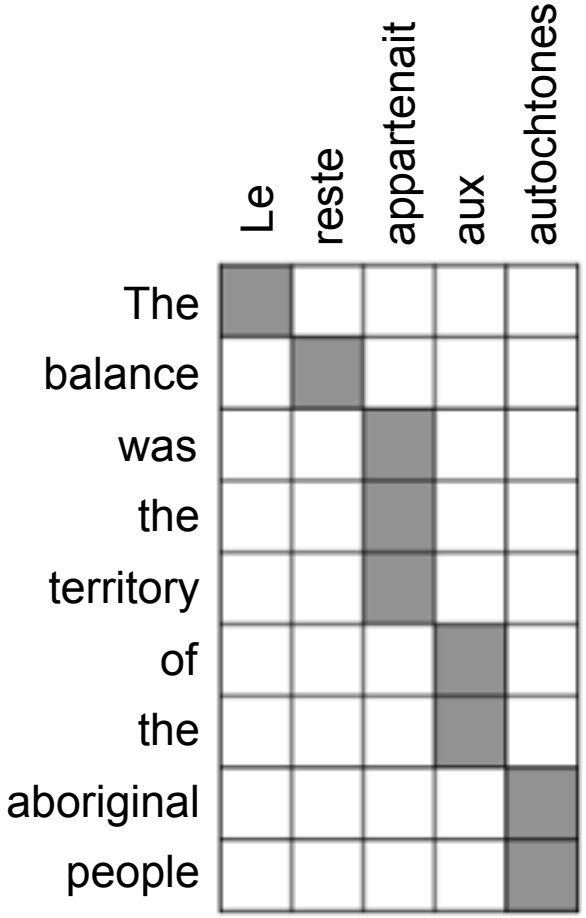


# Statistical Machine Translation

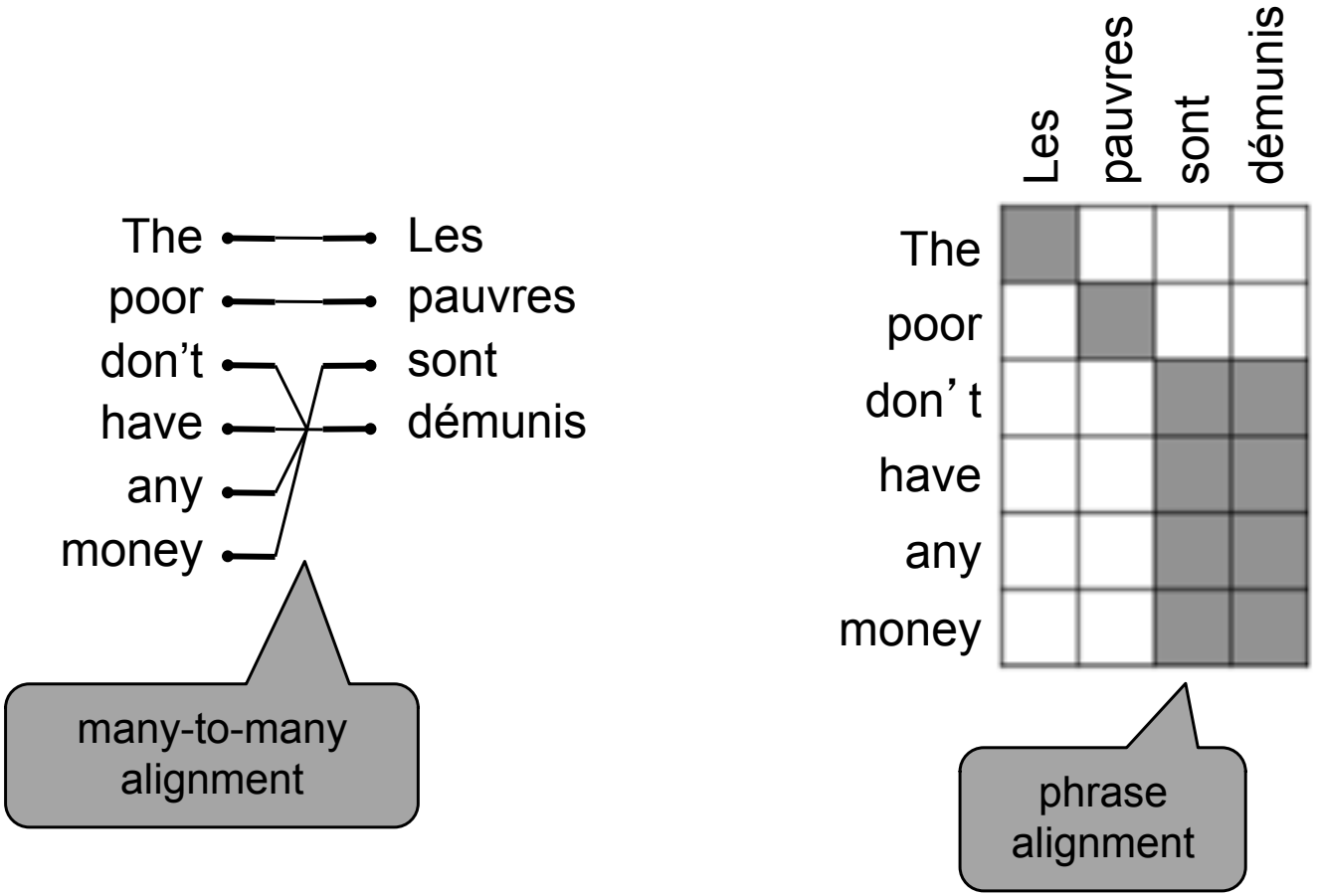
- Many words in the source sentence might align to a single word in the target sentence:



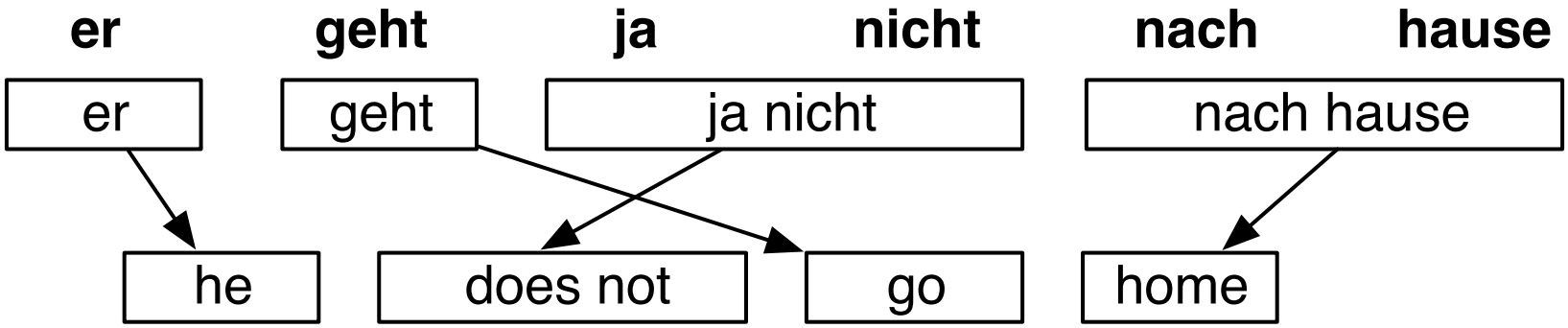
# Statistical Machine Translation

- And finally, many words in the source sentence might align to many words in the target sentence:



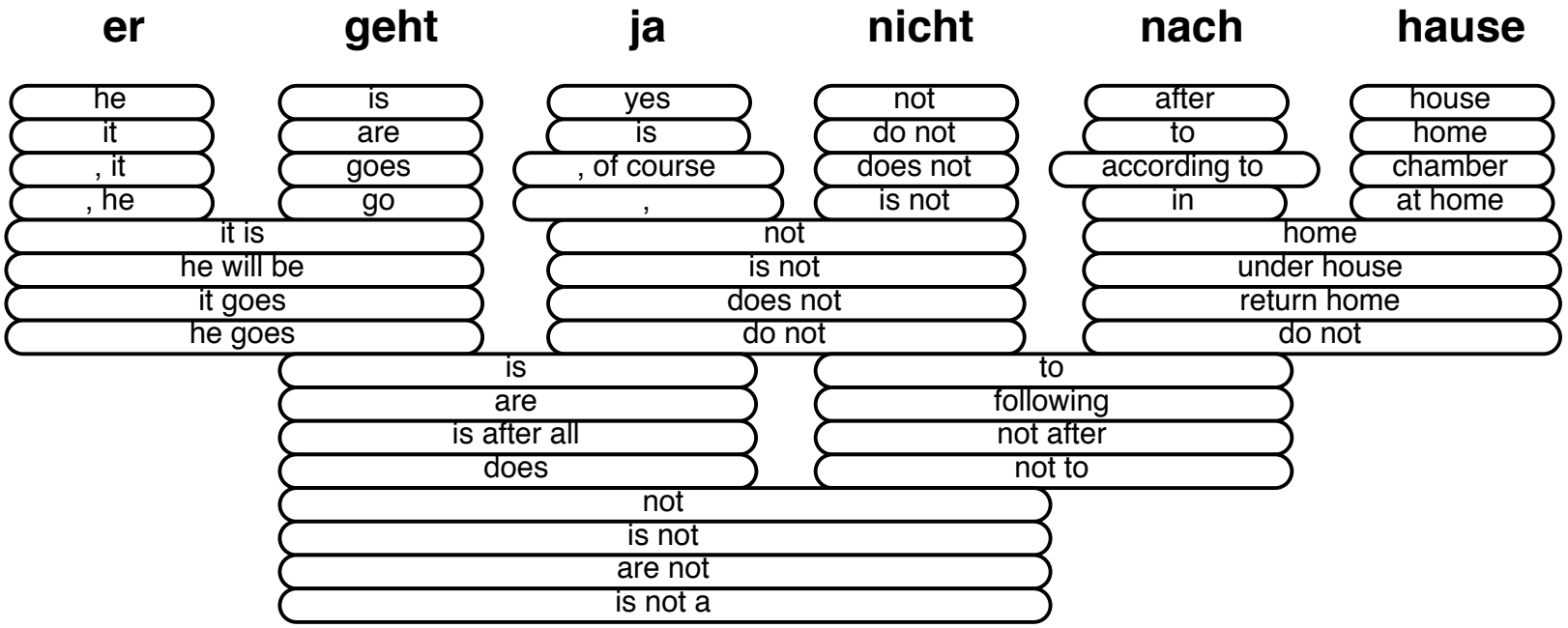
# Statistical Machine Translation

- After learning the word and phrase alignments, the model also needs to figure out the reordering, esp. important in language pairs with very different orders!



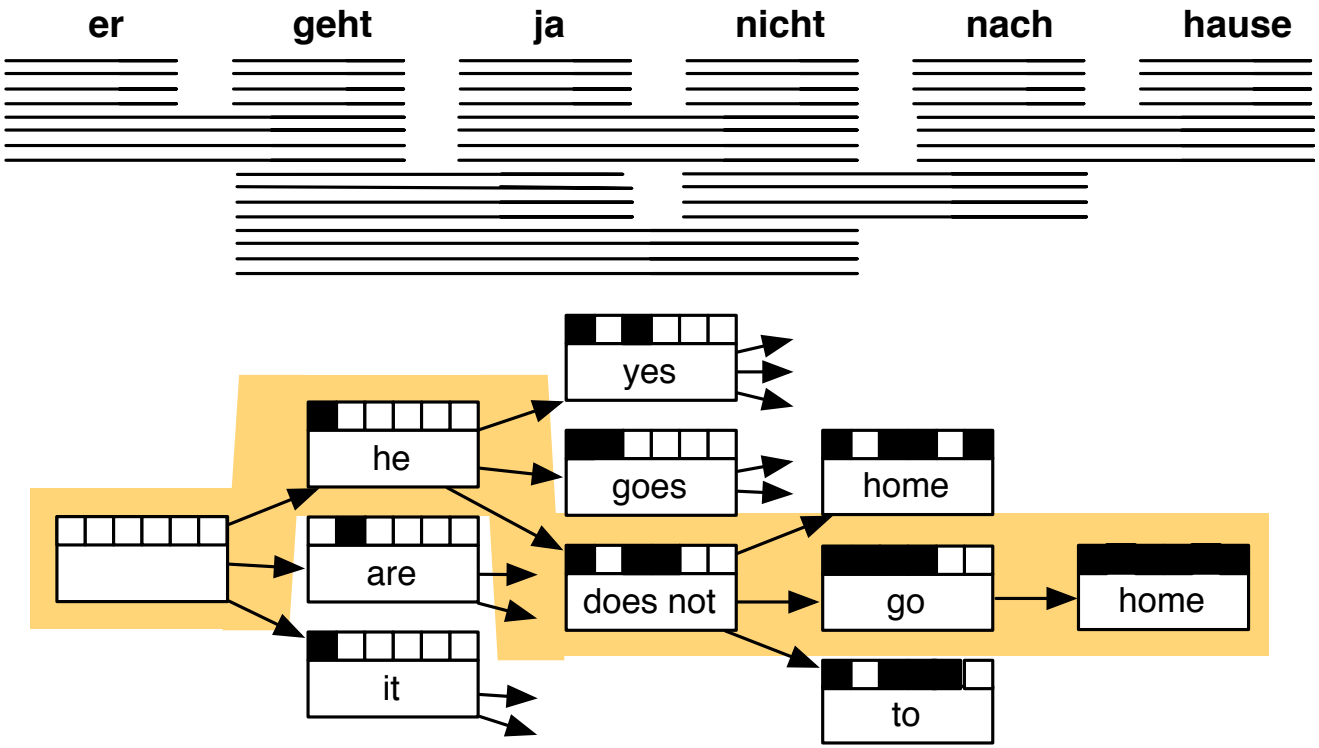
# Statistical Machine Translation

- After many steps, you get the large ‘phrase table’. Each phrase in the source language can have many possible translations in the target language, and hence the search space can be combinatorially large!



# Statistical Machine Translation

- Finally, you decode this hard search problem to find the best translation, e.g., using beam search on the several combinatorial paths through this phrase table (and also include the language model p(e) to rerank)



# Next Week

- IBM Alignment Model Details
- HMM Alignment Model
- Syntactic Models
- Neural Machine Translation (NMT)