Natural Language Processing (Lecture 2)

MLSS 2018

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

Mohit Bansal

(some slides adapted/borrowed from courses by Dan Klein, Richard Socher, Chris Manning, JurafskyMartin-SLP3 book, others)

Question Answering
(& Compositional Semantics 2: Logical forms and Semantic Parsing)

Knowledge Base Q&A (Semantic Parsing)

- Answering question by mapping it to a query (e.g., based on logical forms) executable on a structured database

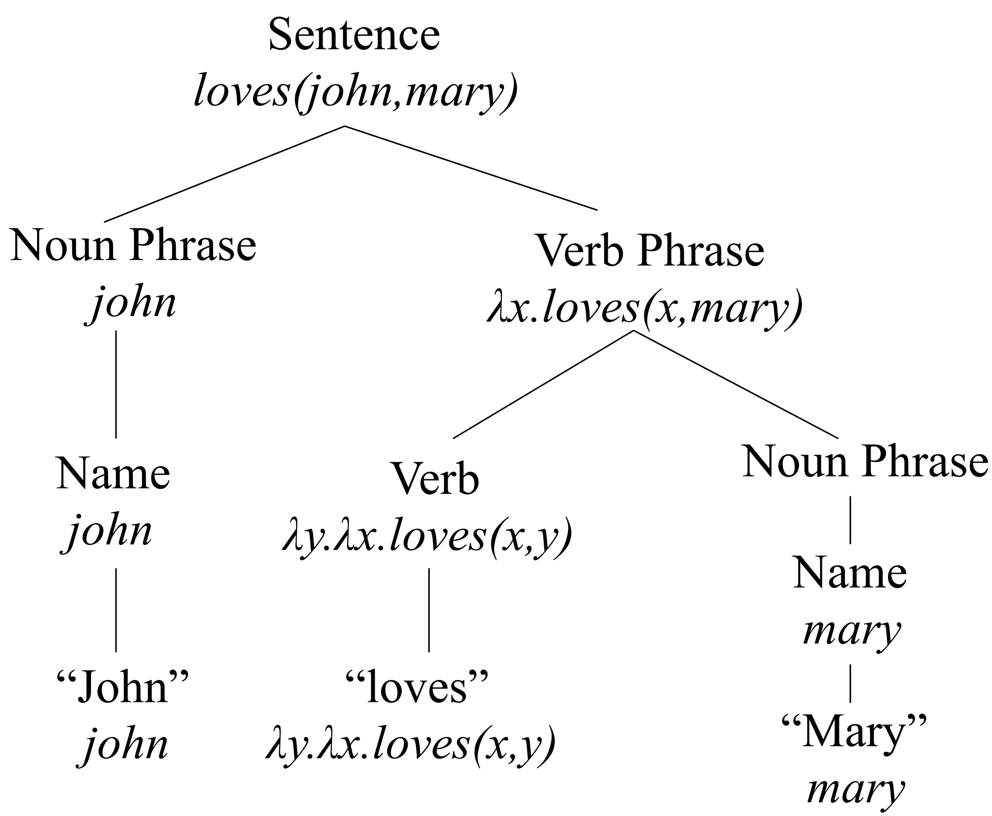
Question	Logical form
When was Ada Lovelace born?	`birth-year (Ada Lovelace, ?x)`
What states border Texas?	λ x.state(x) $\wedge$ borders(x,texas)
What is the largest state	argmax(λx.state(x), λx.size(x))
How many people survived the sinking of the Titanic	`(count (!fb:event.disaster.survivors fb:en.sinking_of_the_titanic))`

Figure 28.7 Sample logical forms produced by a semantic parser for question answering. These range from simple relations like `birth-year`, or relations normalized to databases like Freebase, to full predicate calculus.

[Jurafsky-SLP3]

Semantic Parsing Overview

- Parsing with logic (booleans, individuals, functions) and lambda forms

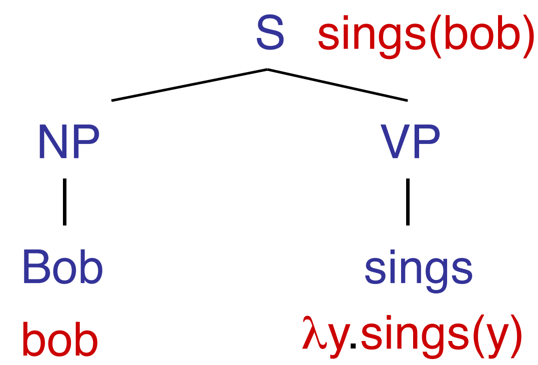


[Wong and Mooney, 2007; Zettlemoyer and Collins, 2007; Poon and Domingos, 2009; Artzi and Zettlemoyer, 2011, 2013; Kwiatkowski et al., 2013; Cai and Yates, 2013; Berant et al., 2013; Poon 2013; Berant and Liang, 2014; Iyyer et al., 2014]

Truth-Conditional Semantics

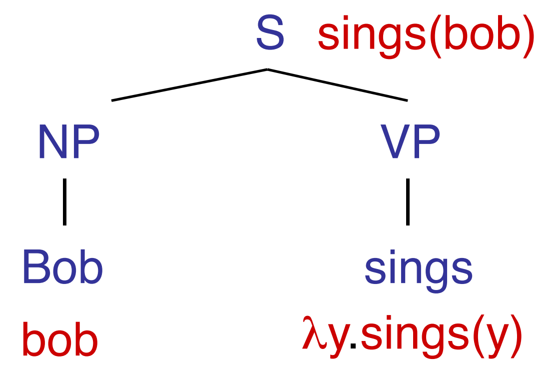
- Examples like “Bob sings”
- Logical translation of this will be something like: sings(bob)
- Types on these translations are entities (e) and truth-values (t), e.g.:

  bob: e
  sings(bob): t



Truth-Conditional Semantics

- For verbs (and verb phrases), sings combines with bob to produce sings(bob)



- In general, we use lambda-calculus (λ-calculus), i.e., a notation for functions whose arguments have not yet been filled/resolved/satisfied

- $\lambda x.\text{sings}(x)$

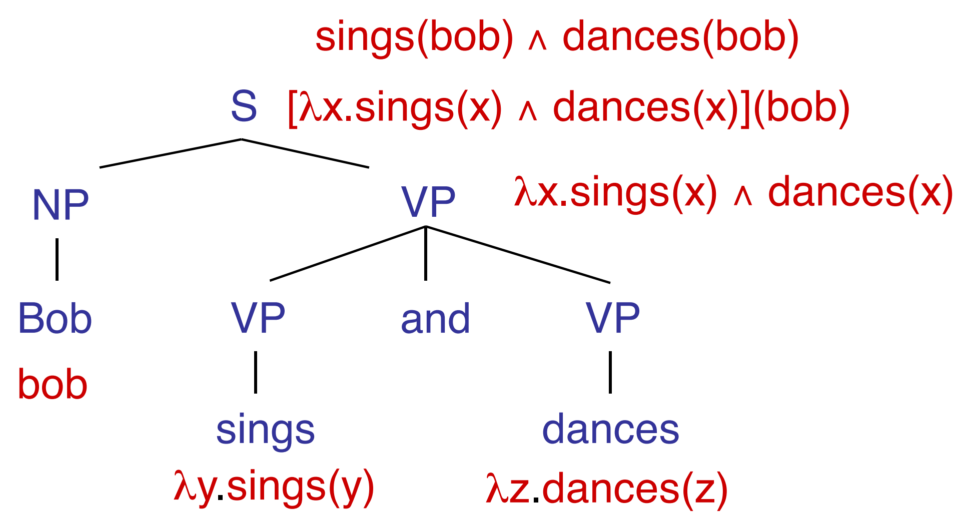
- This is a 'predicate', i.e., a function which take an entity (type e) and produces a truth value (type t), denoted as $e \rightarrow t$

Compositional Semantics 2: Logical forms

- Now after we have these meanings for words, we want to combine them into meaning for phrases and sentences
- For this, we associate a combination rule with each grammar rule of the parse tree, e.g.:

  S: $\beta(\alpha) \rightarrow$ NP: α VP: β (*function application*)

  VP: $\lambda x \,.\, \alpha(x) \wedge \beta(x) \rightarrow$ VP: α and: $\varnothing$ VP: β (*intersection*)



Weak Supervision

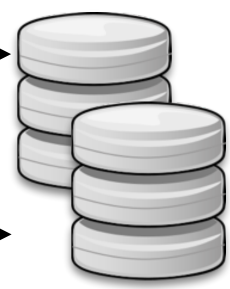
- Instead of relying on sentence—logical form pairs as training data, we can learn from query-answer pairs
- Logical forms are latent, and we can check which one gets the correct answer on being executed against a knowledge base (KB)

What is the largest state that borders Texas?

$$New\ Mexico$$

$$argmax(\lambda x.state(x) \wedge border(x, TX), \lambda y.size(y)) \rightarrow$$

$$argmax(\lambda x.river(x) \wedge in(x, TX), \lambda y.size(y)) \rightarrow$$



$\rightarrow New\ Mexico$ ✓

$\rightarrow Rio\ Grande$ ✗

[Clarke et al. 2010; Liang et al. 2011]

Weak Supervision

- Learning from Instruction-Demonstration Pairs

at the chair, move forward three steps past the sofa

Some examples from other domains:

- Sentences and labeled game states [Goldwasser and Roth 2011]
- Sentences and sets of physical objects [Matuszek et al. 2012]

[Chen and Mooney 2011; Kim and Mooney 2012;Artzi and Zettlemoyer 2013b]

Neural Semantic Parsing

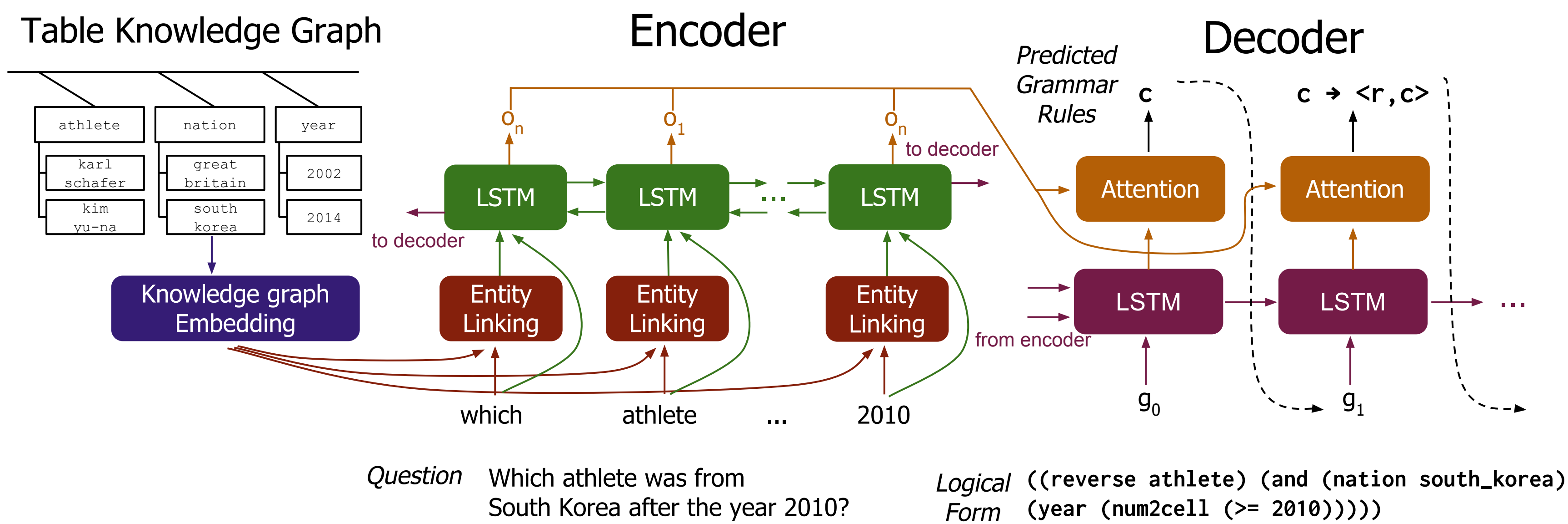


Figure 1: Overview of our semantic parsing model. The encoder performs entity embedding and linking before encoding the question with a bidirectional LSTM. The decoder predicts a sequence of grammar rules that generate a well-typed logical form.

[Krishnamurthy et al., 2017]

Neural AMR Parsing

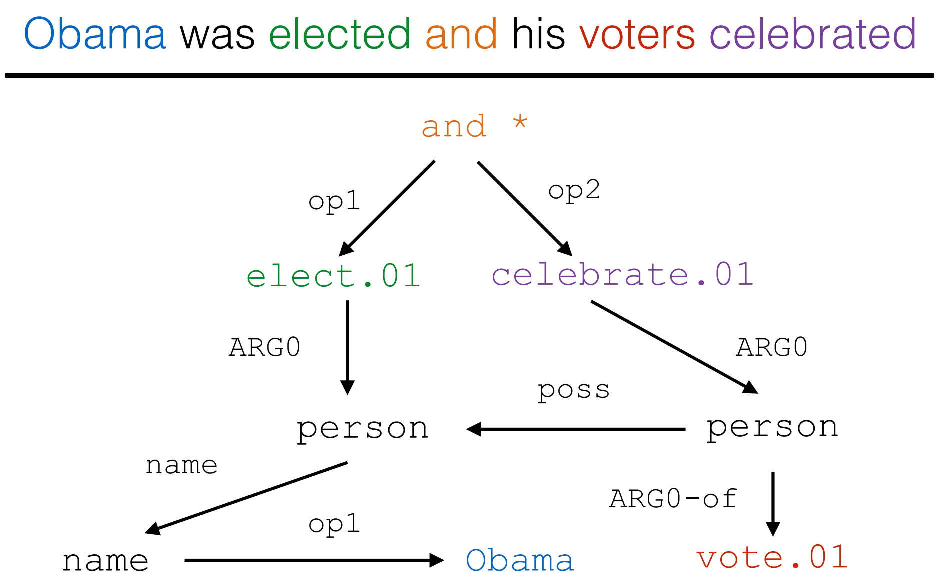
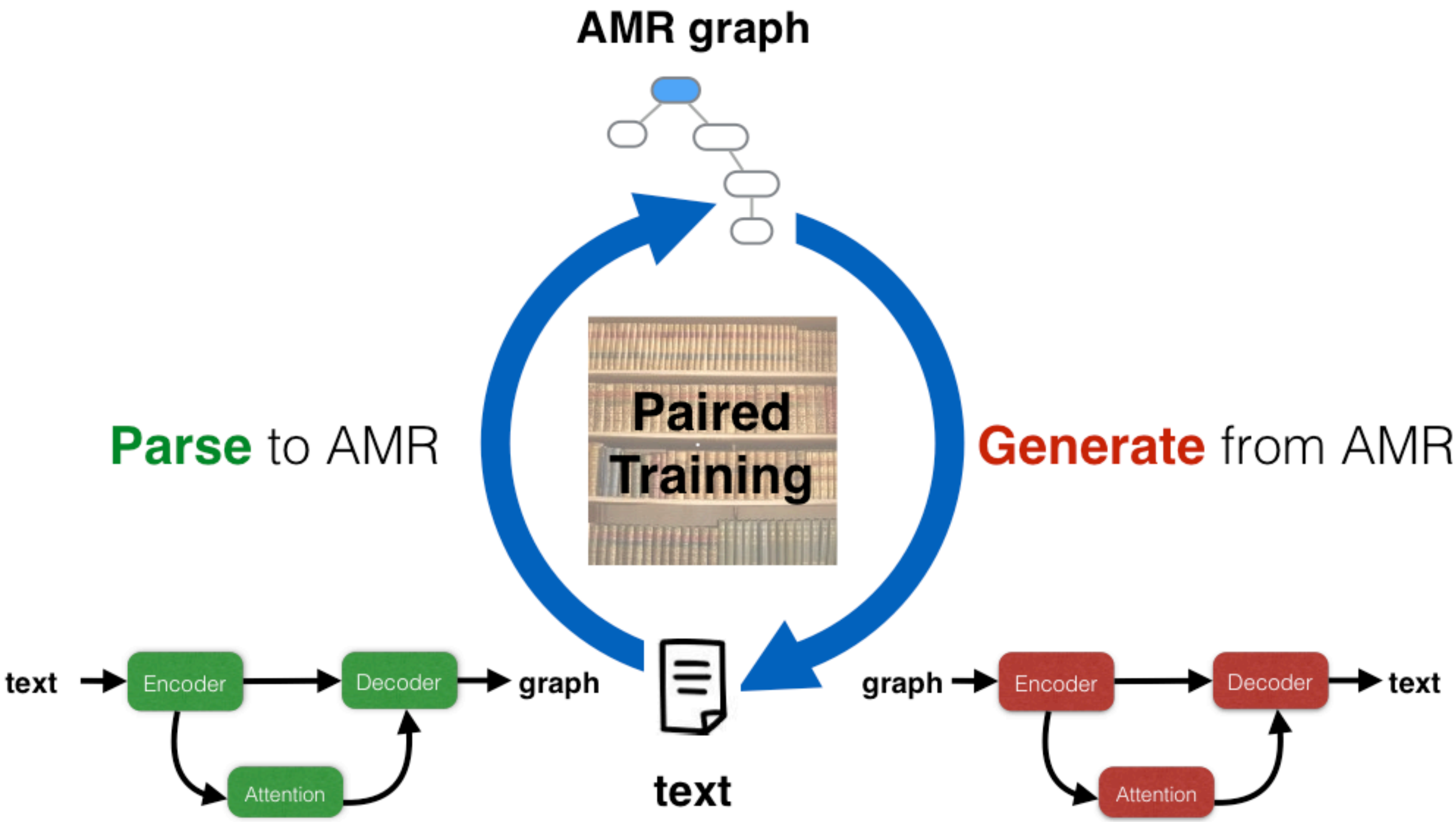


Figure 1: An example sentence and its corresponding Abstract Meaning Representation (AMR). AMR encodes semantic dependencies between entities mentioned in the sentence, such as "Obama" being the "arg0" of the verb "elected".

[Konstas et al., 2017]

Neural AMR Parsing



[Konstas et al., 2017]

IR-based Question Answering

- Information-Retrieval approaches to Q&A: question processing, passage retrieval, answer processing

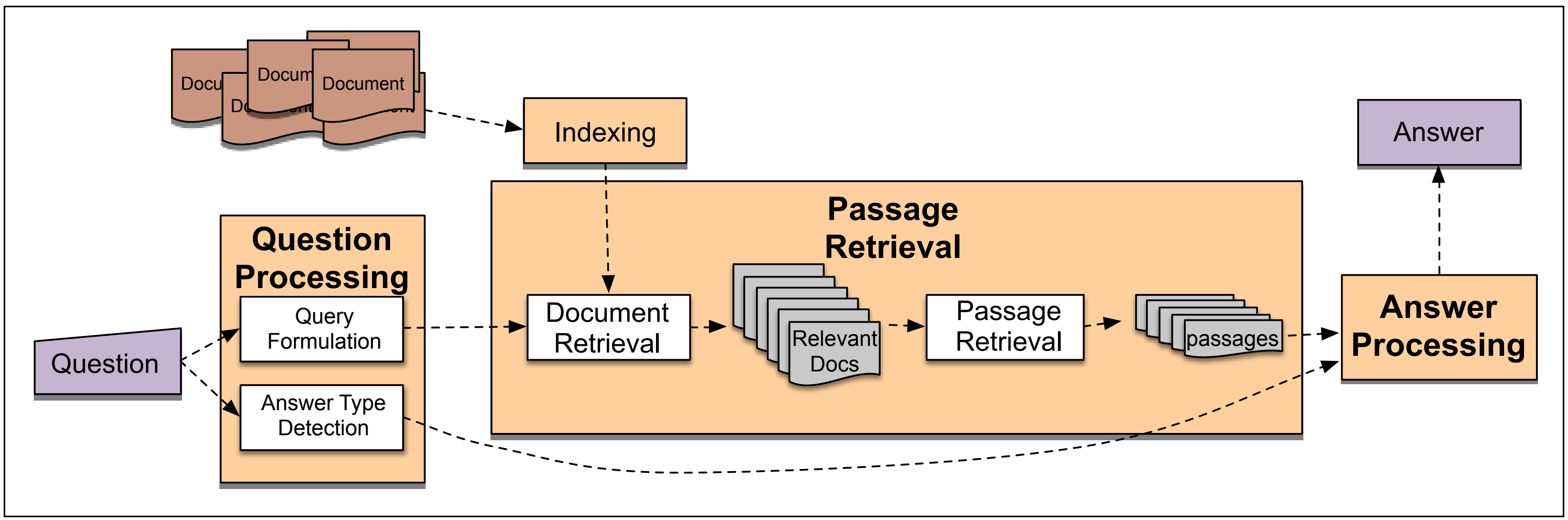


Figure 28.2 IR-based factoid question answering has three stages: question processing, passage retrieval, and answer processing.

[Jurafsky-SLP3]

IR-based Question Answering

- Information-Retrieval approaches to Q&A: question processing, passage retrieval, answer processing

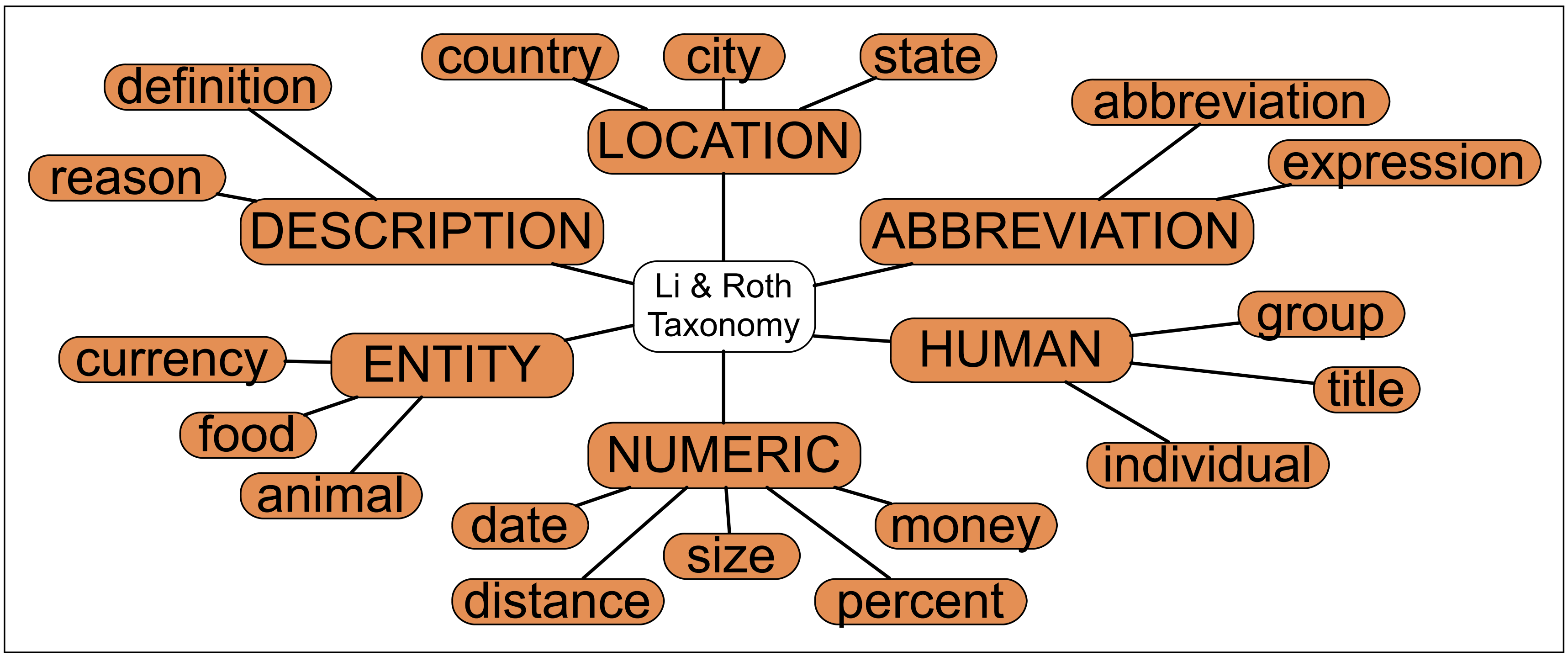


Figure 28.3 A subset of the Li and Roth (2005) answer types.

[Jurafsky-SLP3]

IR-based Question Answering

- Next came a large-scale, open-domain IE system like IBM Watson

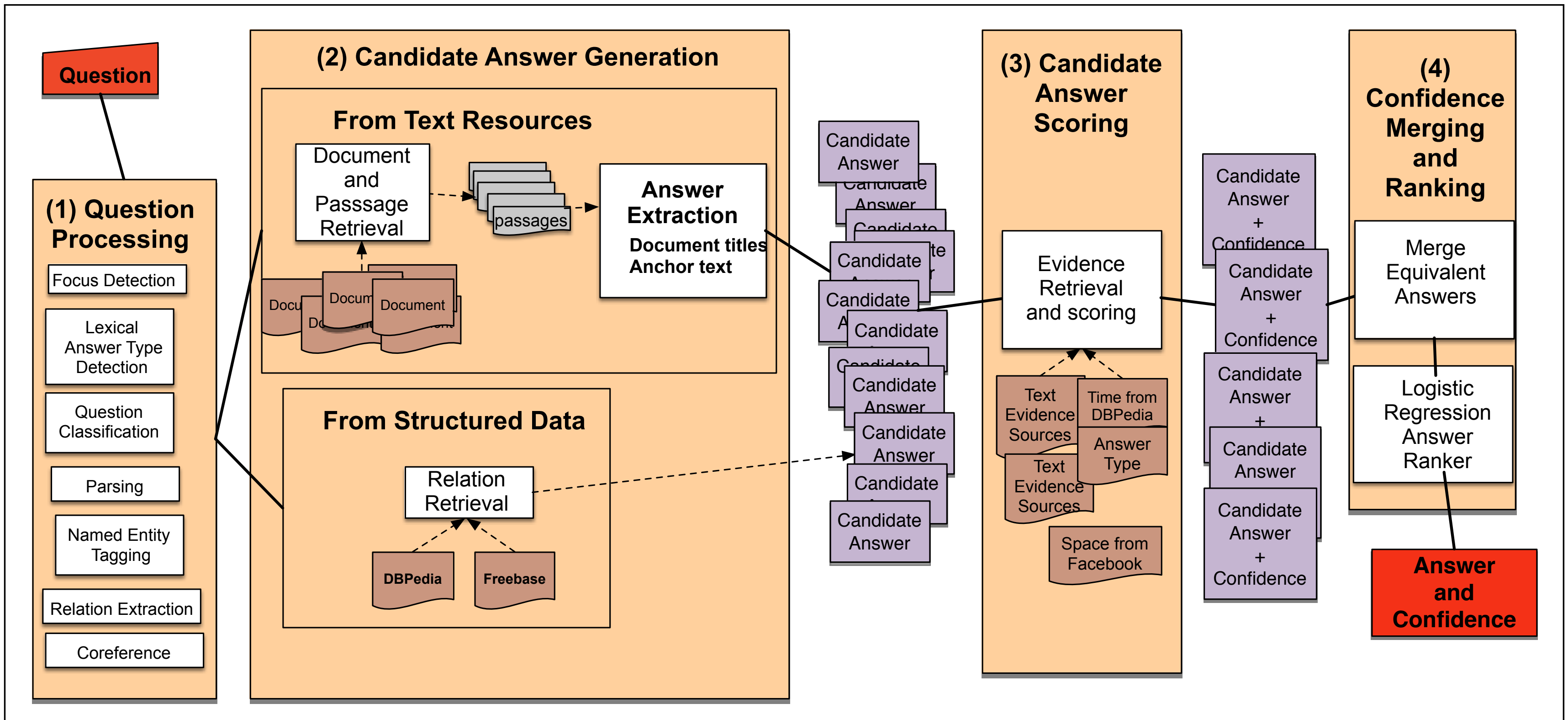
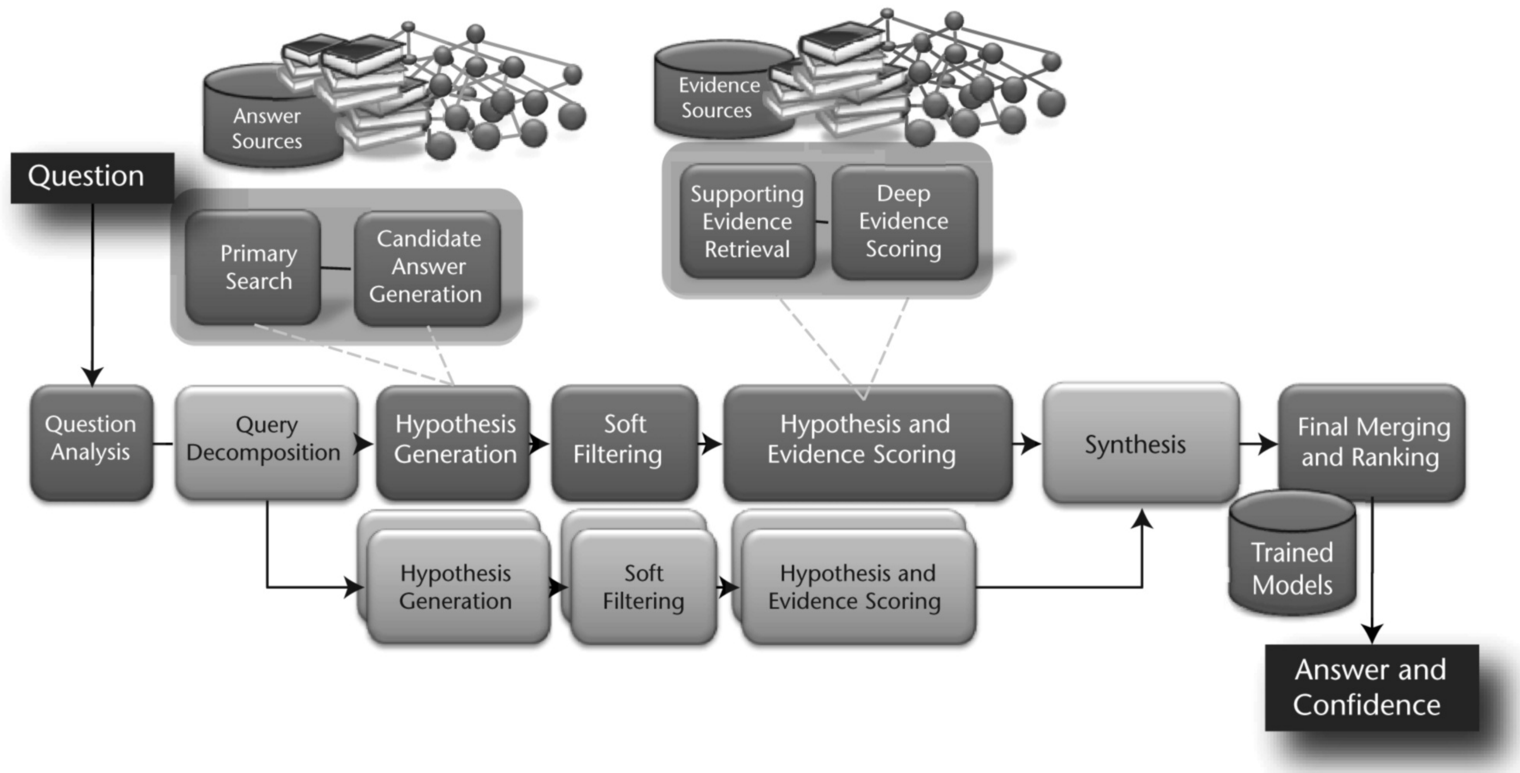


Figure 28.9 The 4 broad stages of Watson QA: (1) Question Processing, (2) Candidate Answer Generation, (3) Candidate Answer Scoring, and (4) Answer Merging and Confidence Scoring.

[Jurafsky-SLP3]

IR-based Question Answering

- Next came a large-scale, open-domain IE system like IBM Watson



[Ferrucci et al., 2010]

Passage-based Q&A (CNN/DailyMail)

	Original Version	Anonymised Version
Context		
	The BBC producer allegedly struck by Jeremy Clarkson will not press charges against the "Top Gear" host, his lawyer said Friday. Clarkson, who hosted one of the most-watched television shows in the world, was dropped by the BBC Wednesday after an internal investigation by the British broadcaster found he had subjected producer Oisin Tymon "to an unprovoked physical and verbal attack." ...	the *ent381* producer allegedly struck by *ent212* will not press charges against the " *ent153* " host , his lawyer said friday . *ent212* , who hosted one of the most - watched television shows in the world , was dropped by the *ent381* wednesday after an internal investigation by the *ent180* broadcaster found he had subjected producer *ent193* " to an unprovoked physical and verbal attack . " ...
Query		
	Producer **X** will not press charges against Jeremy Clarkson, his lawyer says.	producer **X** will not press charges against *ent212* , his lawyer says .
Answer		
	Oisin Tymon	*ent193*

[Hermann et al. 2015]

SQuAd Dataset (100K Manually-Labeled)

- Based on manual annotation from Mturk on Wiki articles, as opposed to cloze/fill-in-the-blank on summaries, etc.; large size (100K+)
- Answer is a span in the document:

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under **gravity**. The main forms of precipitation include drizzle, rain, sleet, snow, **graupel** and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals **within a cloud**. Short, intense periods of rain in scattered locations are called "showers".

What causes precipitation to fall?
gravity

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
graupel

Where do water droplets collide with ice crystals to form precipitation?
within a cloud

[Rajpurkar et al. 2016]

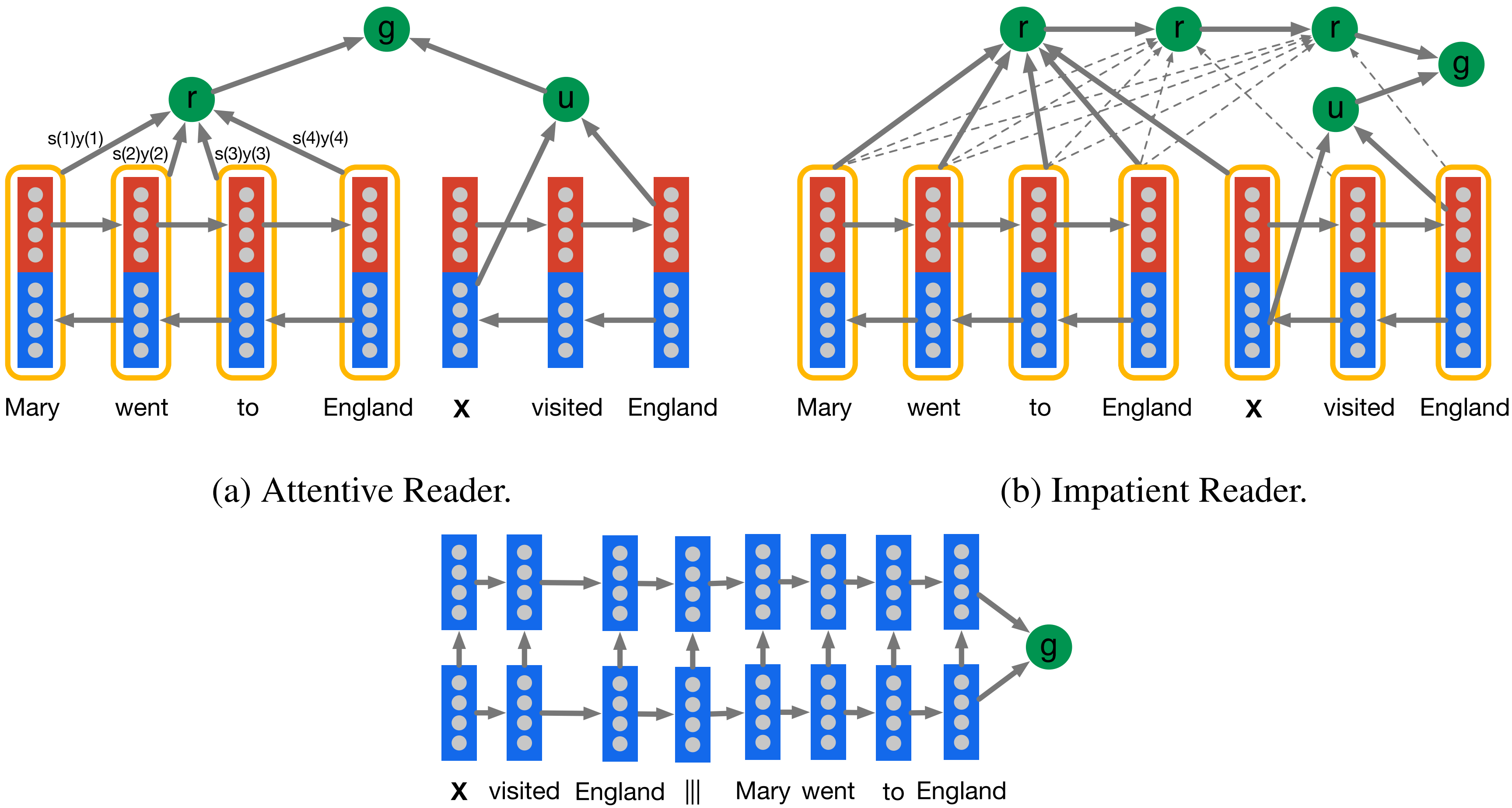
SQuAd Dataset (100K Manually-Labeled)

Dataset	Question source	Formulation	Size
SQuAD	**crowdsourced**	**RC, spans in passage**	**100K**
MCTest (Richardson et al., 2013)	crowdsourced	RC, multiple choice	2640
Algebra (Kushman et al., 2014)	standardized tests	computation	514
Science (Clark and Etzioni, 2016)	standardized tests	reasoning, multiple choice	855
WikiQA (Yang et al., 2015)	query logs	IR, sentence selection	3047
TREC-QA (Voorhees and Tice, 2000)	query logs + human editor	IR, free form	1479
CNN/Daily Mail (Hermann et al., 2015)	summary + cloze	RC, fill in single entity	1.4M
CBT (Hill et al., 2015)	cloze	RC, fill in single word	688K

Table 1: A survey of several reading comprehension and question answering datasets. SQuAD is much larger than all datasets except the semi-synthetic cloze-style datasets, and it is similar to TREC-QA in the open-endedness of the answers.

[Rajpurkar et al. 2016]

Attentive/Impatient Readers



(a) Attentive Reader.

(b) Impatient Reader.

(c) A two layer Deep LSTM Reader with the question encoded before the document.

[Hermann et al. 2015]

Attentive/Impatient Readers

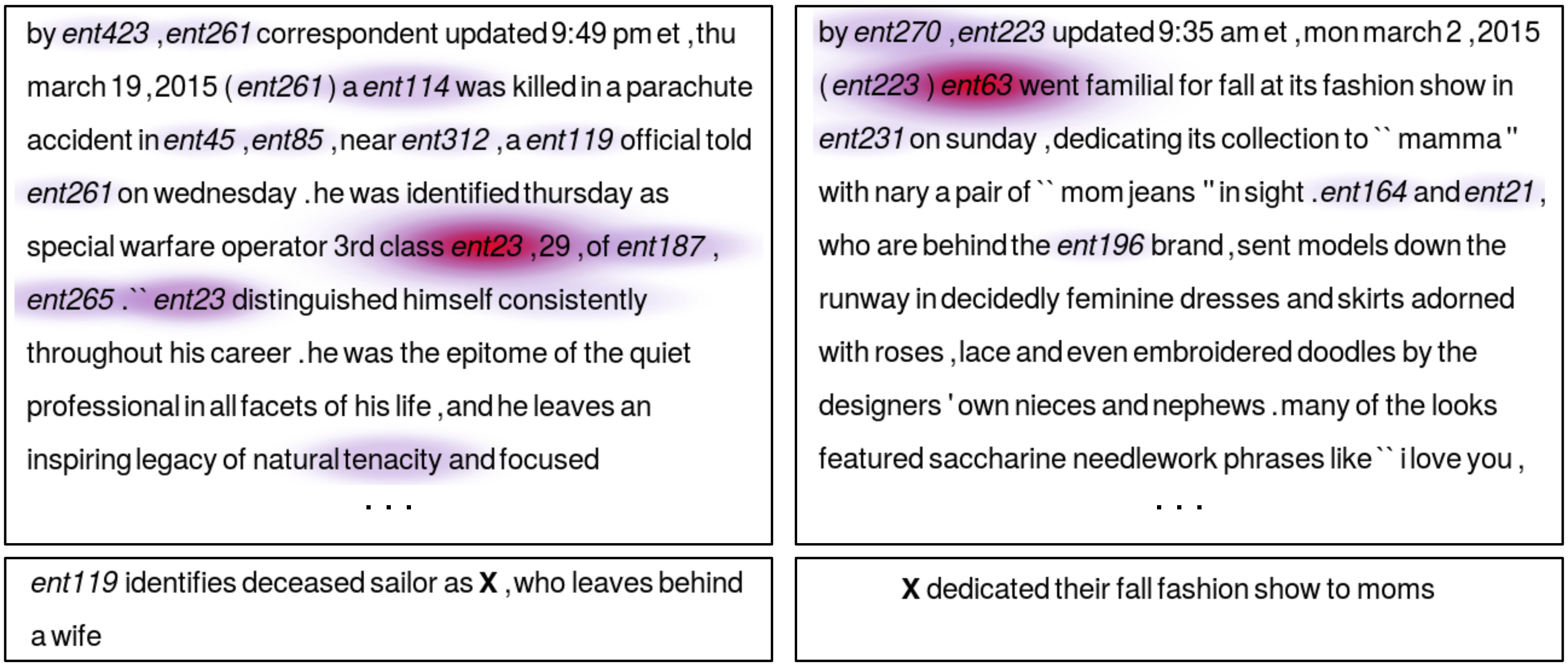
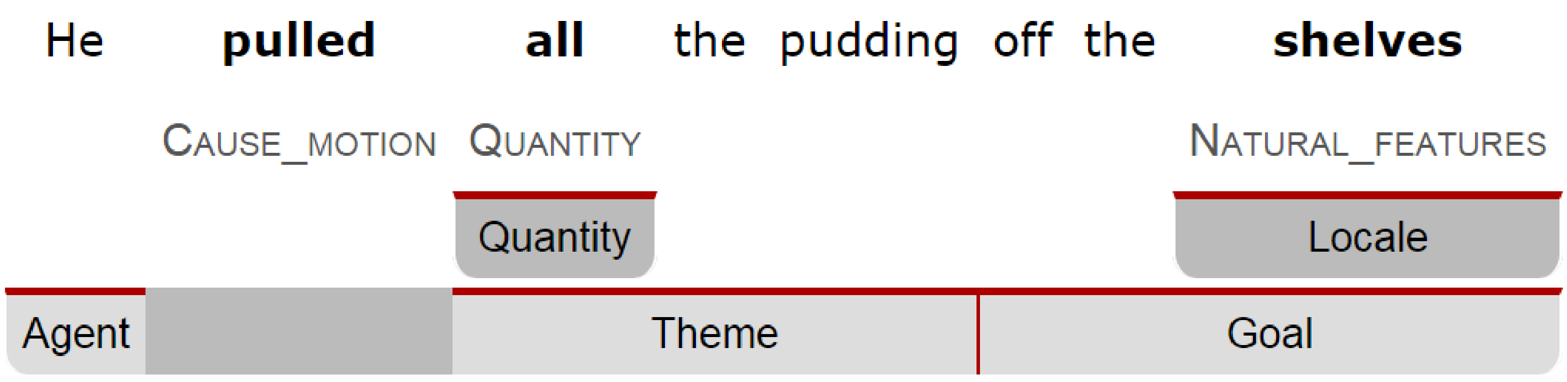


Figure 3: Attention heat maps from the Attentive Reader for two correctly answered validation set queries (the correct answers are *ent23* and *ent63*, respectively). Both examples require significant lexical generalisation and co-reference resolution in order to be answered correctly by a given model.

[Hermann et al. 2015]

Feature-based Model

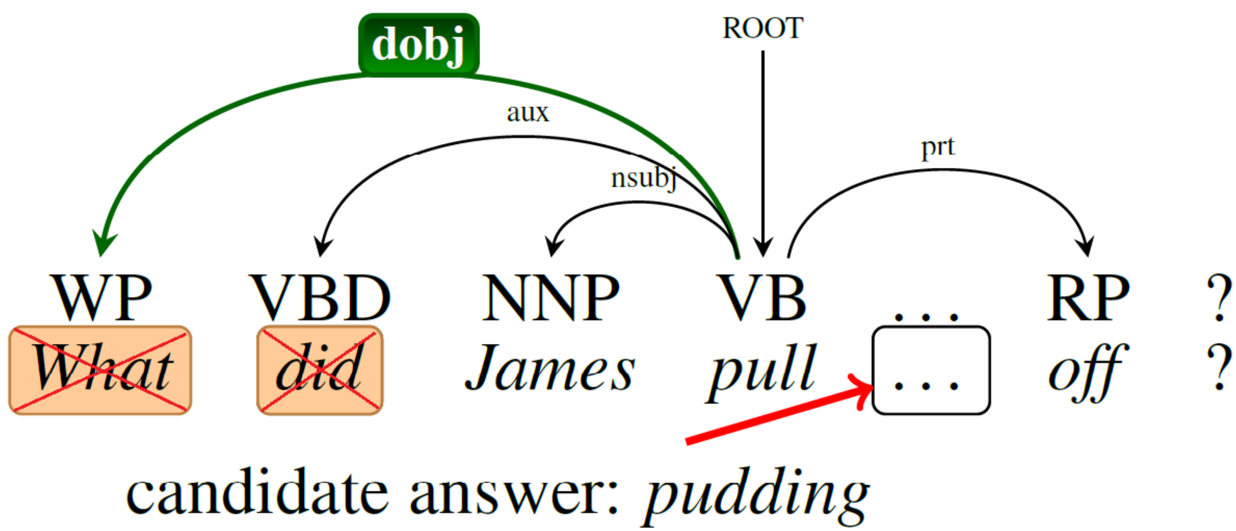
- Weighted word overlap between the bag of words constructed from the question/answer and in the window (and their word embedding versions)
- Minimal distance between two word occurrences in the passage that are also contained in the question/answer pair
- Frame semantics (predicates, frames evoked, and predicted argument labels) match between passage sentence and question+answer



[Wang et al. 2015]

Feature-based Model

- Syntactic dependencies match between passage sentence and ques+ans converted to statement
- Extra features computed after coreference resolution of pronouns/nominals to map to their entity clusters



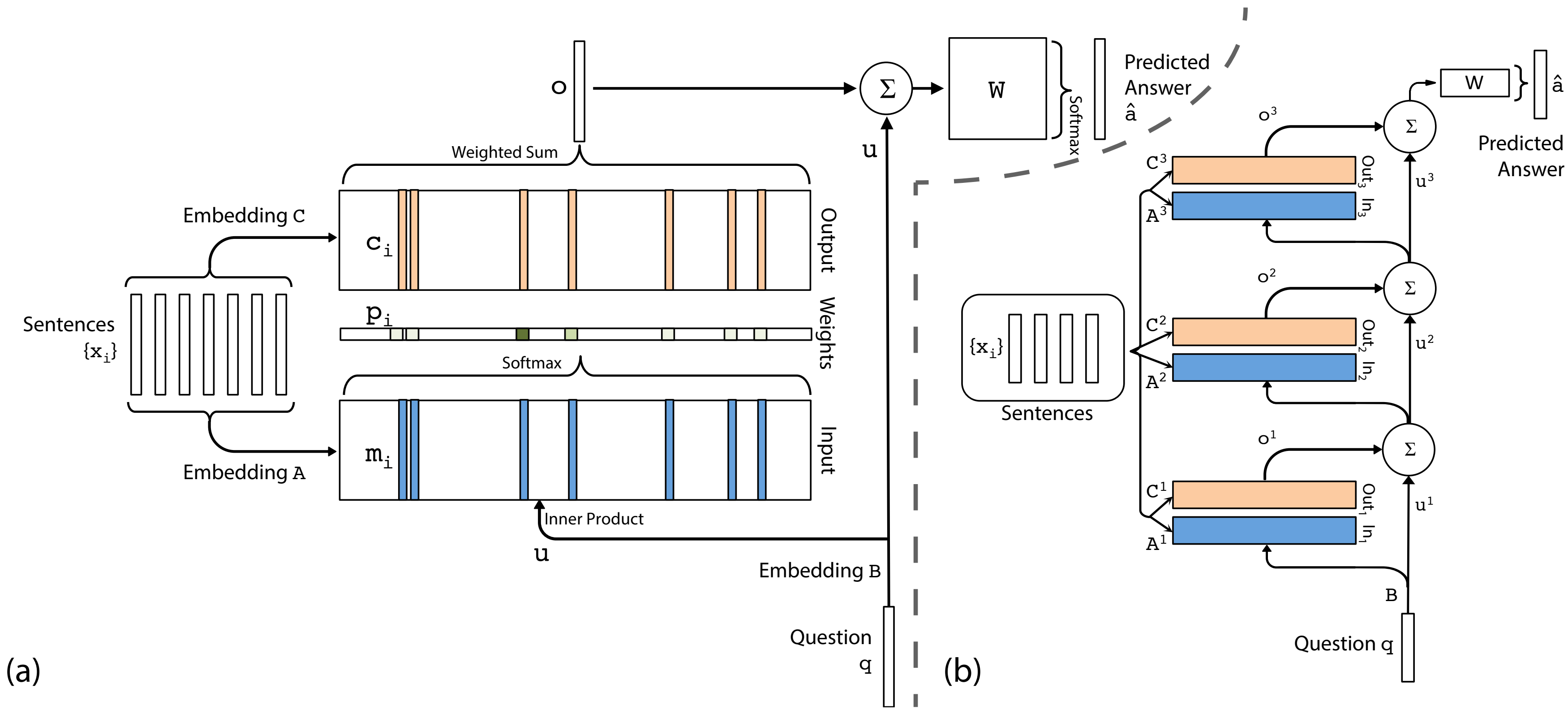
[Wang et al. 2015]

Multi-Hop Memory Models

- Several questions need multi-hop (e.g., path or count-based) reasoning to answer
- Memory models perform multiple passes over the text to collect the multiple evidence pieces
- Some example models:
  - End-to-End Memory Networks
  - Dynamic Memory Networks
  - Gated Attention Readers
  - MAC Cell

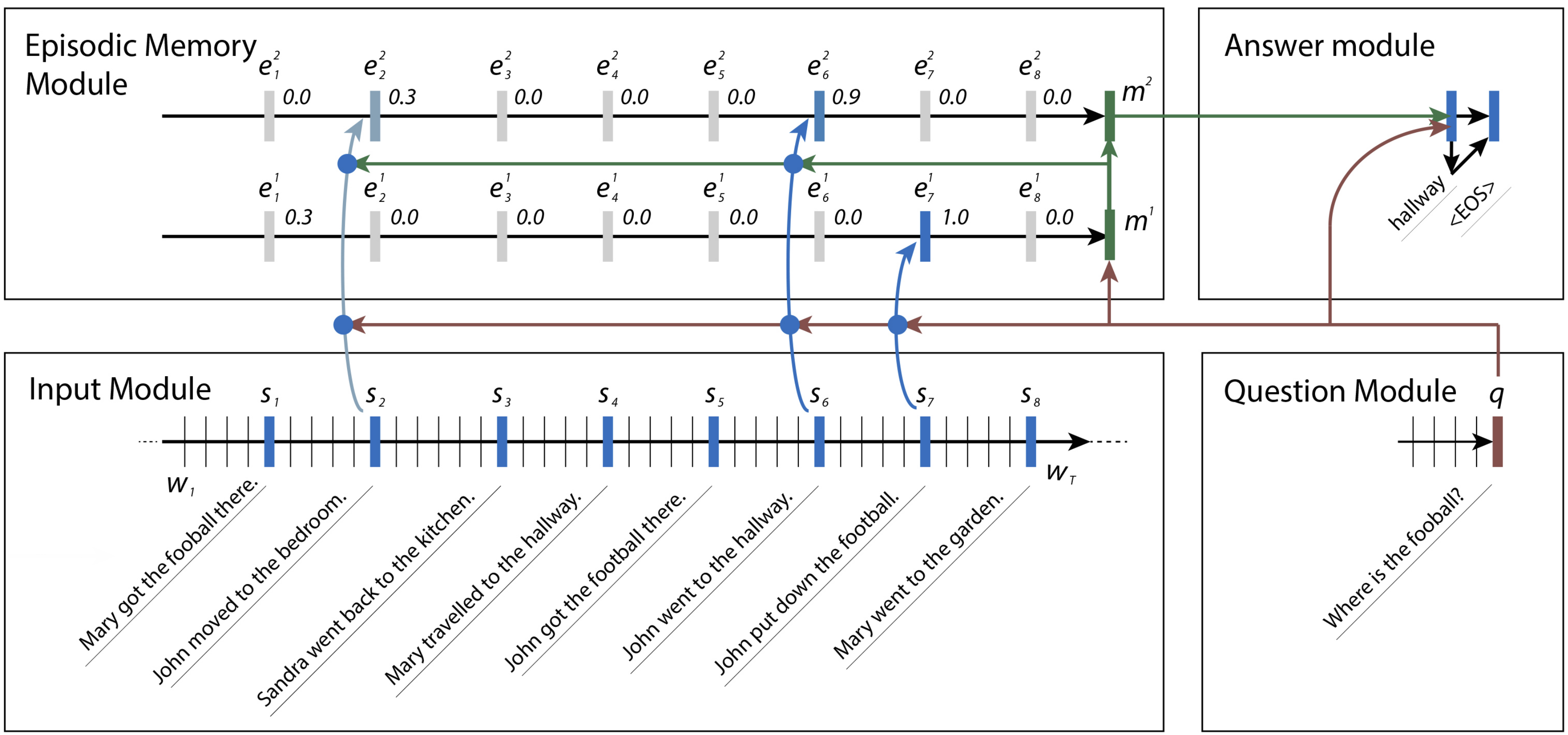
[Weston et al., 2014; Sukhbaatar et al., 2015; Kumar et al., 2016; Dhingra et al. 2017]

End-to-End Memory Networks



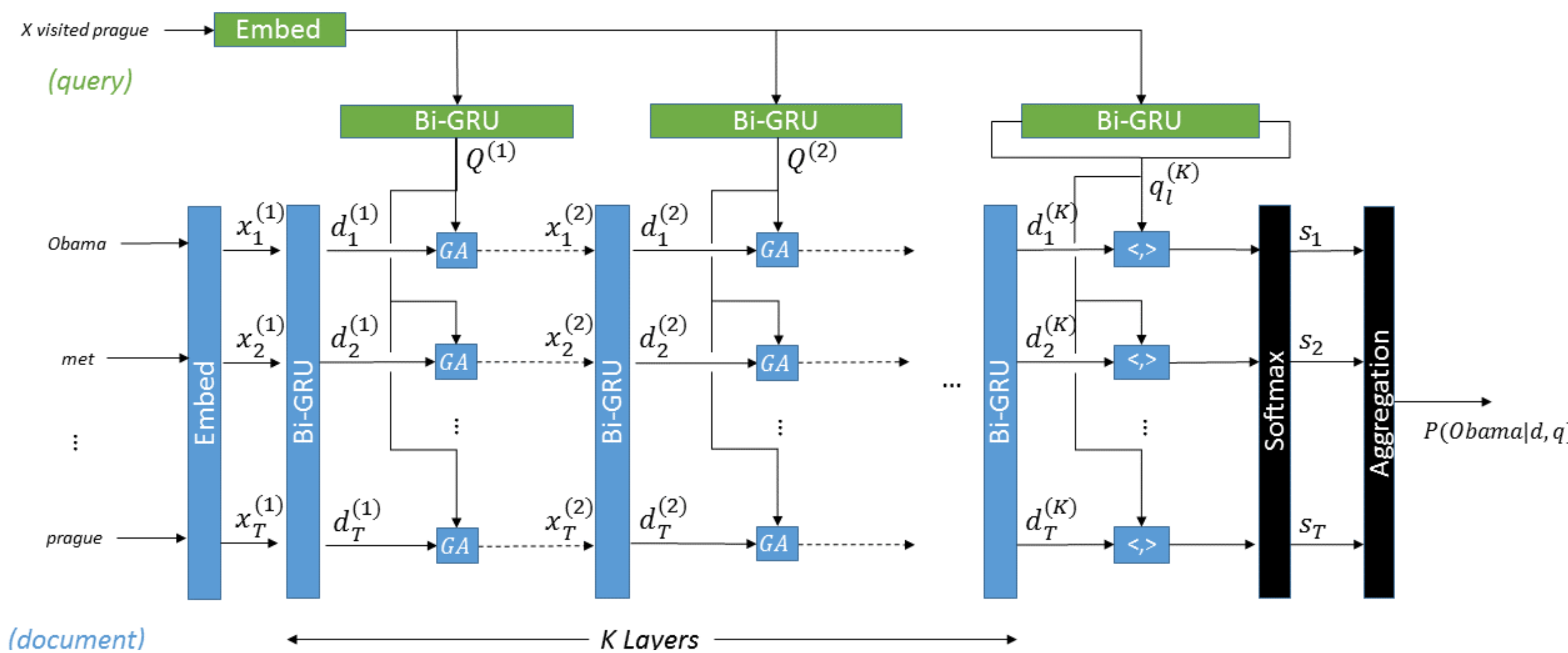
[Sukhbaatar et al., 2015]

Dynamic Memory Networks



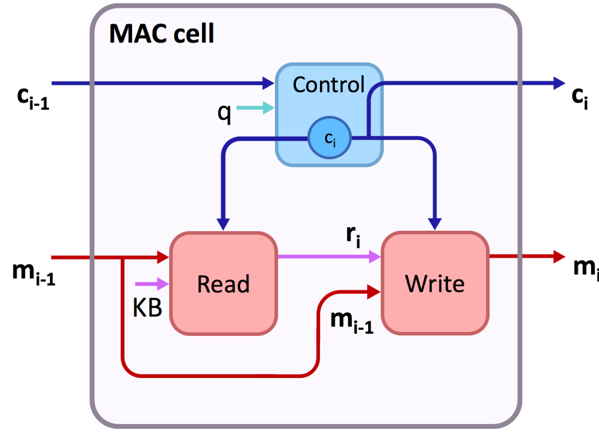
[Kumar et al., 2016]

Gated Attention Readers



[Dhingra et al. 2017]

MAC Cell



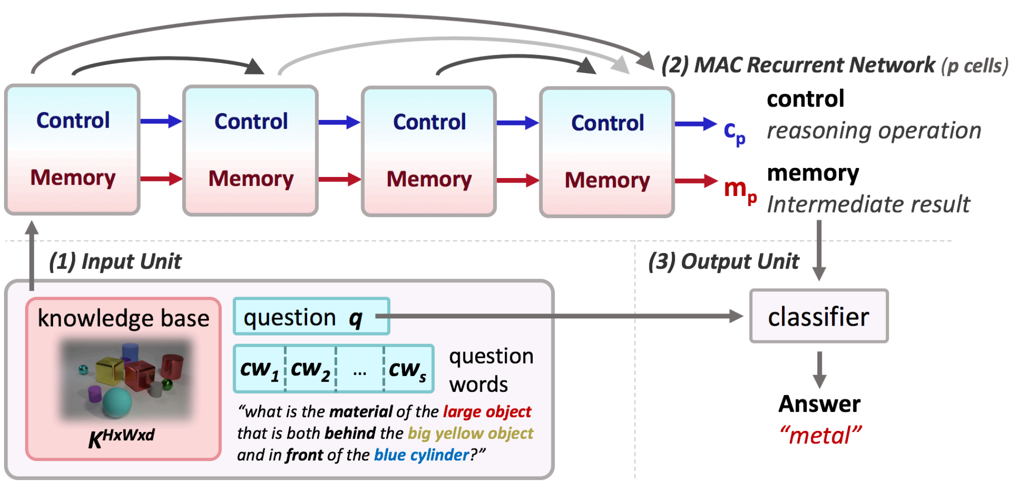


Figure 2: Model Overview. The MAC network consists of an input unit, a core recurrent network and an output unit. (1) The input unit transforms the raw image and question into distributed vector representations. (2) The core recurrent network reasons sequentially over the question by decomposing it into a series of operations (*control*) that retrieve information from the image (knowledge base) and aggregate the results into a recurrent *memory*. (3) The output classifier computes the final answer using the question and the final memory state.

[Hudson & Manning, 2018]

Facebook bAbI Tasks

Task 1: Single Supporting Fact
Mary went to the bathroom.
John moved to the hallway.
Mary travelled to the office.
Where is Mary? A:office

Task 2: Two Supporting Facts
John is in the playground.
John picked up the football.
Bob went to the kitchen.
Where is the football? A:playground

Task 3: Three Supporting Facts
John picked up the apple.
John went to the office.
John went to the kitchen.
John dropped the apple.
Where was the apple before the kitchen? A:office

Task 4: Two Argument Relations
The office is north of the bedroom.
The bedroom is north of the bathroom.
The kitchen is west of the garden.
What is north of the bedroom? A: office
What is the bedroom north of? A: bathroom

Task 5: Three Argument Relations
Mary gave the cake to Fred.
Fred gave the cake to Bill.
Jeff was given the milk by Bill.
Who gave the cake to Fred? A: Mary
Who did Fred give the cake to? A: Bill

Task 6: Yes/No Questions
John moved to the playground.
Daniel went to the bathroom.
John went back to the hallway.
Is John in the playground? A:no
Is Daniel in the bathroom? A:yes

Task 7: Counting
Daniel picked up the football.
Daniel dropped the football.
Daniel got the milk.
Daniel took the apple.
How many objects is Daniel holding? A: two

Task 8: Lists/Sets
Daniel picks up the football.
Daniel drops the newspaper.
Daniel picks up the milk.
John took the apple.
What is Daniel holding? milk, football

Task 9: Simple Negation
Sandra travelled to the office.
Fred is no longer in the office.
Is Fred in the office? A:no
Is Sandra in the office? A:yes

Task 10: Indefinite Knowledge
John is either in the classroom or the playground.
Sandra is in the garden.
Is John in the classroom? A:maybe
Is John in the office? A:no

[Weston et al. 2016]

Facebook bAbI Tasks

Task 11: Basic Coreference
Daniel was in the kitchen.
Then he went to the studio.
Sandra was in the office.
Where is Daniel? A:studio

Task 12: Conjunction
Mary and Jeff went to the kitchen.
Then Jeff went to the park.
Where is Mary? A: kitchen
Where is Jeff? A: park

Task 13: Compound Coreference
Daniel and Sandra journeyed to the office.
Then they went to the garden.
Sandra and John travelled to the kitchen.
After that they moved to the hallway.
Where is Daniel? A: garden

Task 14: Time Reasoning
In the afternoon Julie went to the park.
Yesterday Julie was at school.
Julie went to the cinema this evening.
Where did Julie go after the park? A:cinema
Where was Julie before the park? A:school

Task 15: Basic Deduction
Sheep are afraid of wolves.
Cats are afraid of dogs.
Mice are afraid of cats.
Gertrude is a sheep.
What is Gertrude afraid of? A:wolves

Task 16: Basic Induction
Lily is a swan.
Lily is white.
Bernhard is green.
Greg is a swan.
What color is Greg? A:white

Task 17: Positional Reasoning
The triangle is to the right of the blue square.
The red square is on top of the blue square.
The red sphere is to the right of the blue square.
Is the red sphere to the right of the blue square? A:yes
Is the red square to the left of the triangle? A:yes

Task 18: Size Reasoning
The football fits in the suitcase.
The suitcase fits in the cupboard.
The box is smaller than the football.
Will the box fit in the suitcase? A:yes
Will the cupboard fit in the box? A:no

Task 19: Path Finding
The kitchen is north of the hallway.
The bathroom is west of the bedroom.
The den is east of the hallway.
The office is south of the bedroom.
How do you go from den to kitchen? A: west, north
How do you go from office to bathroom? A: north, west

Task 20: Agent's Motivations
John is hungry.
John goes to the kitchen.
John grabbed the apple there.
Daniel is hungry.
Where does Daniel go? A:kitchen
Why did John go to the kitchen? A:hungry

[Weston et al. 2016]

QAngaroo Dataset

The Hanging Gardens, in **[Mumbai]**, also known as Pherozeshah Mehta Gardens, are terraced gardens … They provide sunset views over the **[Arabian Sea]** …

Mumbai (also known as Bombay, the official name until 1995) is the capital city of the Indian state of Maharashtra. It is the most populous city in **India** …

The **Arabian Sea** is a region of the northern Indian Ocean bounded on the north by **Pakistan** and **Iran**, on the west by northeastern **Somalia** and the Arabian Peninsula, and on the east by **India** …

Q: (Hanging gardens of Mumbai, country, ?)
Options: {Iran, **India**, Pakistan, Somalia, …}

Figure 1: A sample from the WIKIHOP dataset where it is necessary to combine information spread across multiple documents to infer the correct answer.

Leuprolide ... elicited a long-lasting potentiation of excitatory postsynaptic currents… **[GnRH receptor]**-induced synaptic potentiation was blocked … by **[Progonadoliberin-1]**, a specific **[GnRH receptor]** antagonist…

… our research to study the distribution, co-localization of **Urofollitropin** and its receptor[,] and co-localization of **Urofollitropin** and **GnRH receptor**…

Analyses of gene expression demonstrated a dynamic response to the **Progonadoliberin-1** superagonist **Triptorelin**.

Q: (Leuprolide, interacts_with, ?)
Options: {**Triptorelin, Urofollitropin**}

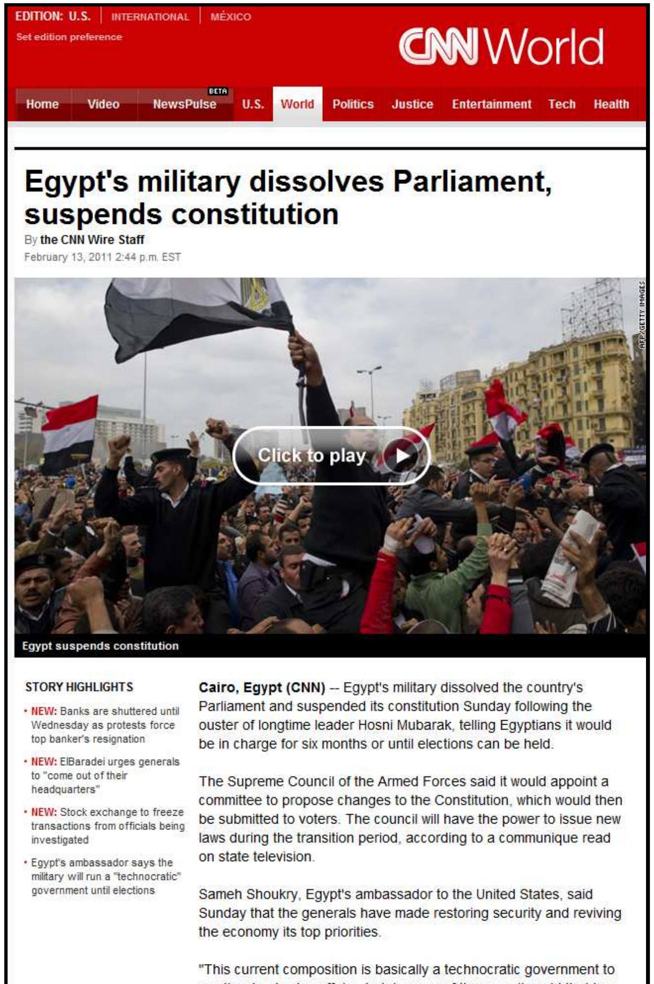
Figure 3: A sample from the MEDHOP dataset.

[Welbl et al. 2018]

Automatic Document Summarization

Single-Document Summarization

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



EDITION: U.S. | INTERNATIONAL | MÉXICO

Set edition preference

CNN World

Home | Video | NewsPulse BETA | U.S. | World | Politics | Justice | Entertainment | Tech | Health

Egypt's military dissolves Parliament, suspends constitution

By **the CNN Wire Staff**

February 13, 2011 2:44 p.m. EST

Egypt suspends constitution

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

Cairo, Egypt (CNN) -- Egypt's military dissolved the country's Parliament and suspended its constitution Sunday following the ouster of longtime leader Hosni Mubarak, telling Egyptians it would be in charge for six months or until elections can be held.

The Supreme Council of the Armed Forces said it would appoint a committee to propose changes to the Constitution, which would then be submitted to voters. The council will have the power to issue new laws during the transition period, according to a communique read on state television.

Sameh Shoukry, Egypt's ambassador to the United States, said Sunday that the generals have made restoring security and reviving the economy its top priorities.

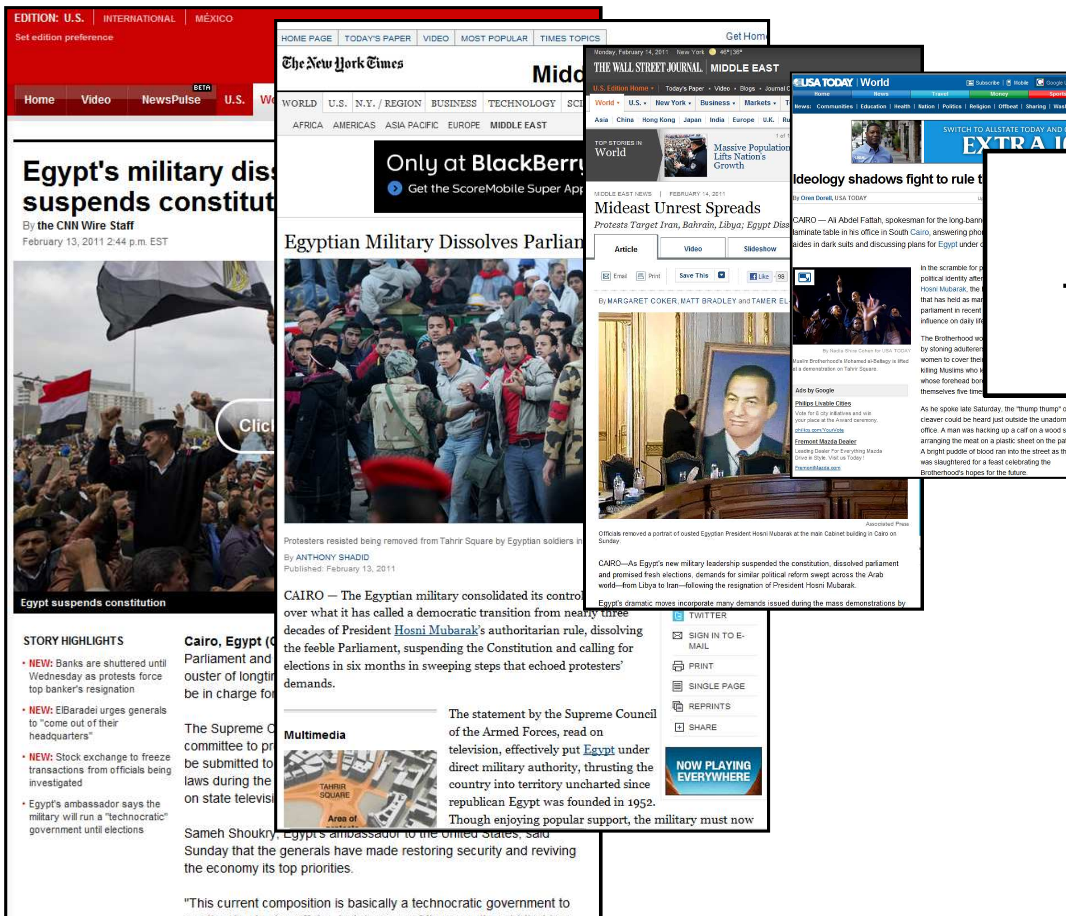
"This current composition is basically a technocratic government to run the day-to-day affairs, to take care of the security void that has

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)



… 27,000+ more

Extractive Summarization

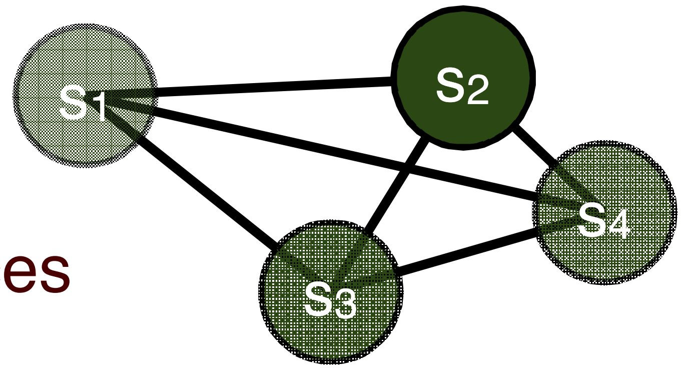
- Directly selecting existing sentences from input document instead of rewriting them

s1	The health care bill is a major test for the Obama administration.
s2	Universal health care is a divisive issue.
s3	President Obama remained calm.
s4	Obama addressed the House on Tuesday.

Graph-based Extractive Summarization

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

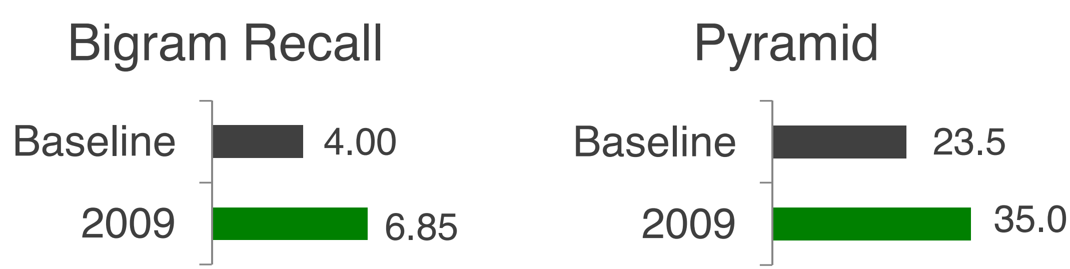
$$\max_{s \in S(D)} \sum_{c \in C(s)} v_c$$

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

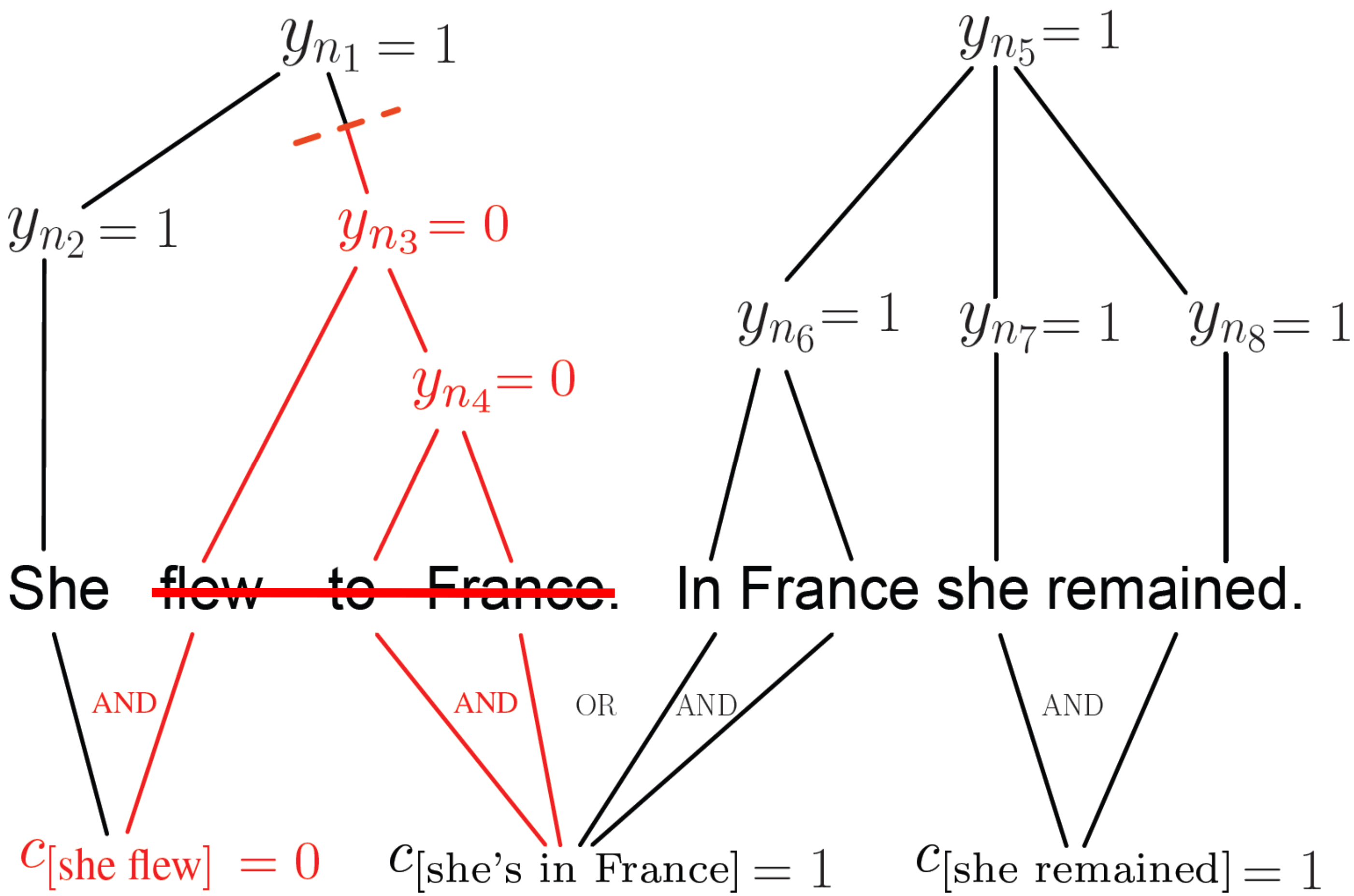
What would a human do?

~~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

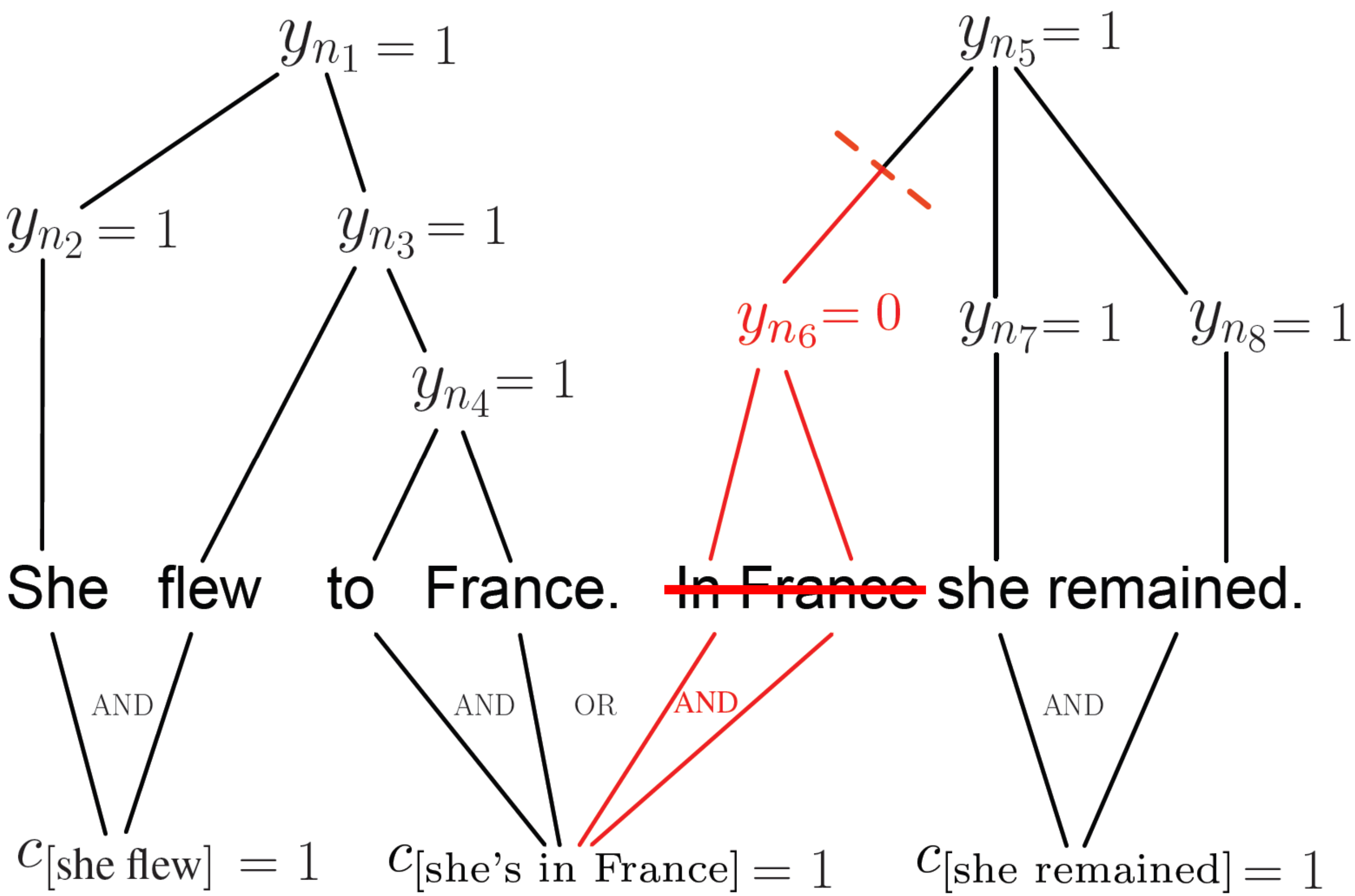
- Model should learn the subtree deletions/cuts that allow compression



[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 $f(c)$

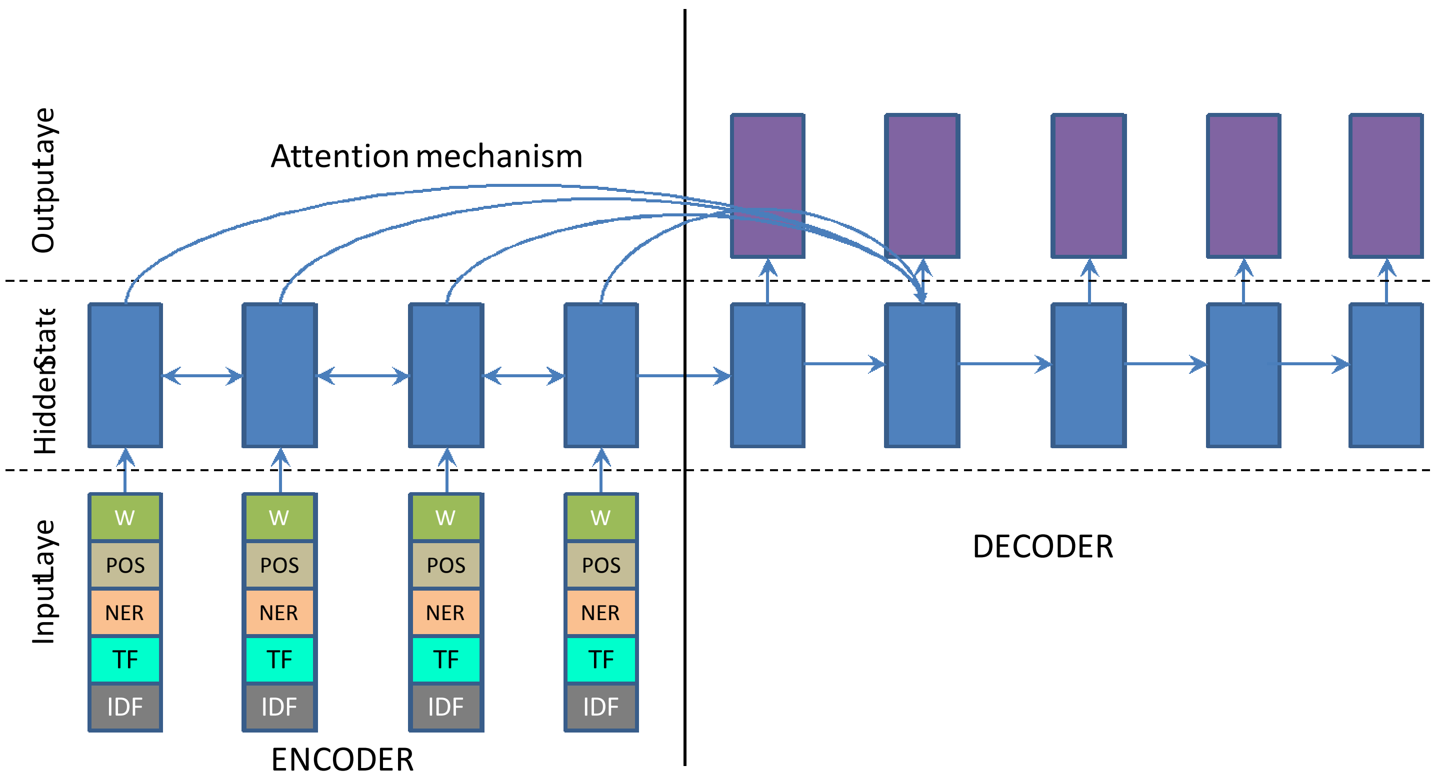
COORD:	Coordinated phrase, four versions: NP, VP, S, SBAR
S-ADJUNCT:	Adjunct to matrix verb, four versions: CC, PP, ADVP, SBAR
REL-C:	Relative clause indicator
ATTR-C:	Attribution clause indicator
ATTR-PP:	PP attribution indicator
TEMP-PP:	Temporal PP indicator
TEMP-NP	Temporal NP indicator
BIAS:	Always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

[Berg-Kirkpatrick et al., 2011]

Neural Abstractive Summarization

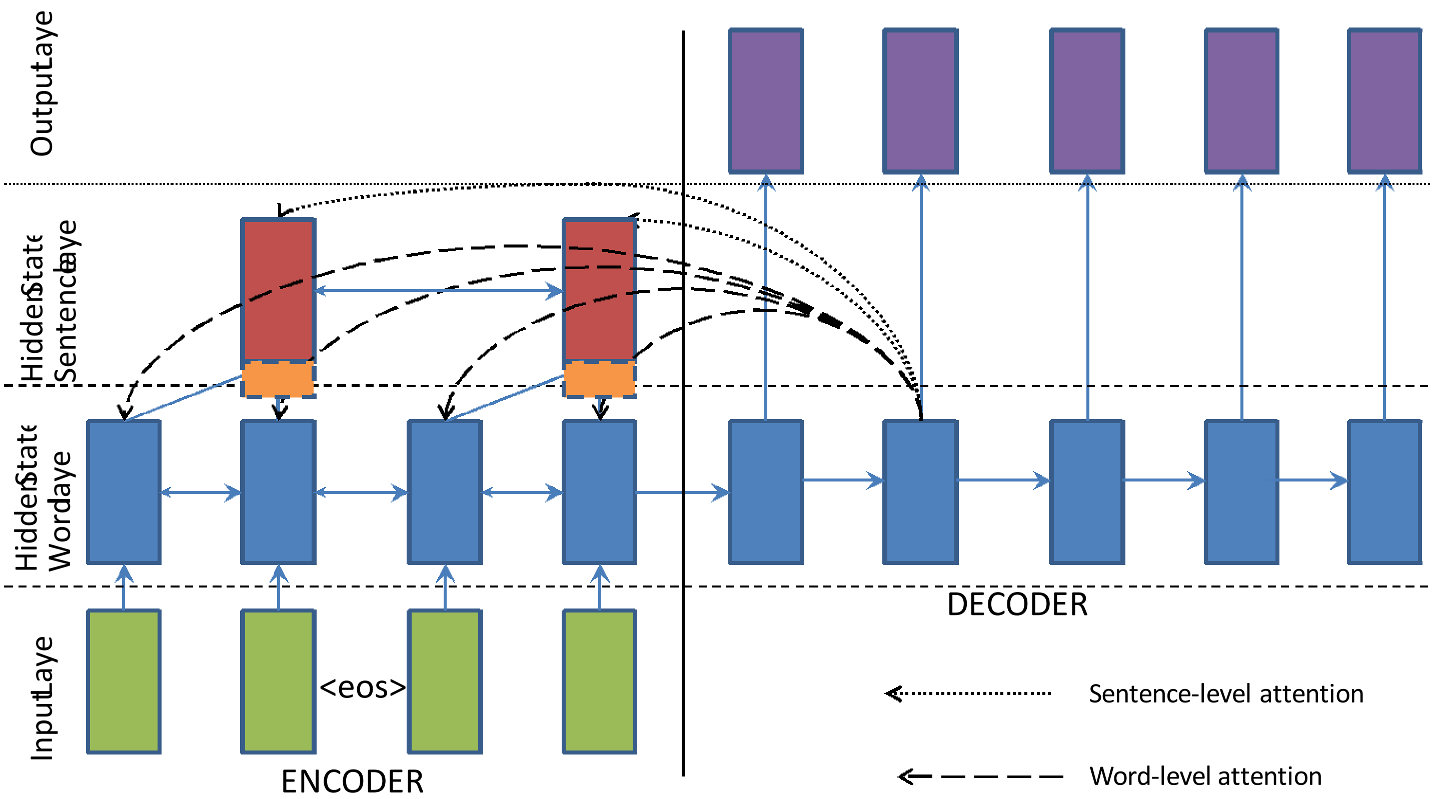
- Mostly based on sequence-to-sequence RNN models
- Later added attention, coverage, pointer/copy, hierarchical encoder/attention, metric rewards RL, etc.
- Examples: Rush et al., 2015; Nallapati et al., 2016; See et al., 2017; Paulus et al., 2017; Chen and Bansal, 2018

Feature-Augmented Encoder-Decoder



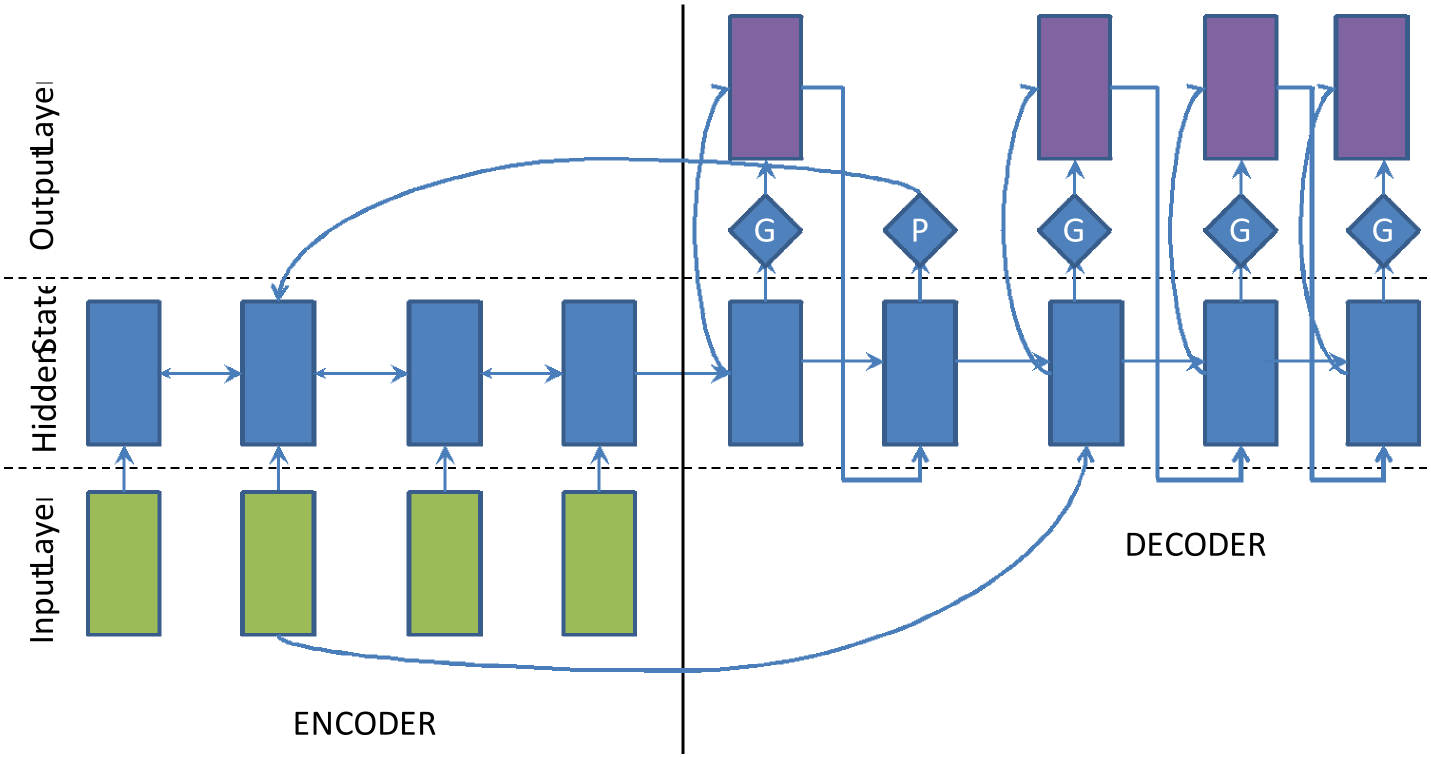
[Nallapati et al., 2016]

Hierarchical Attention



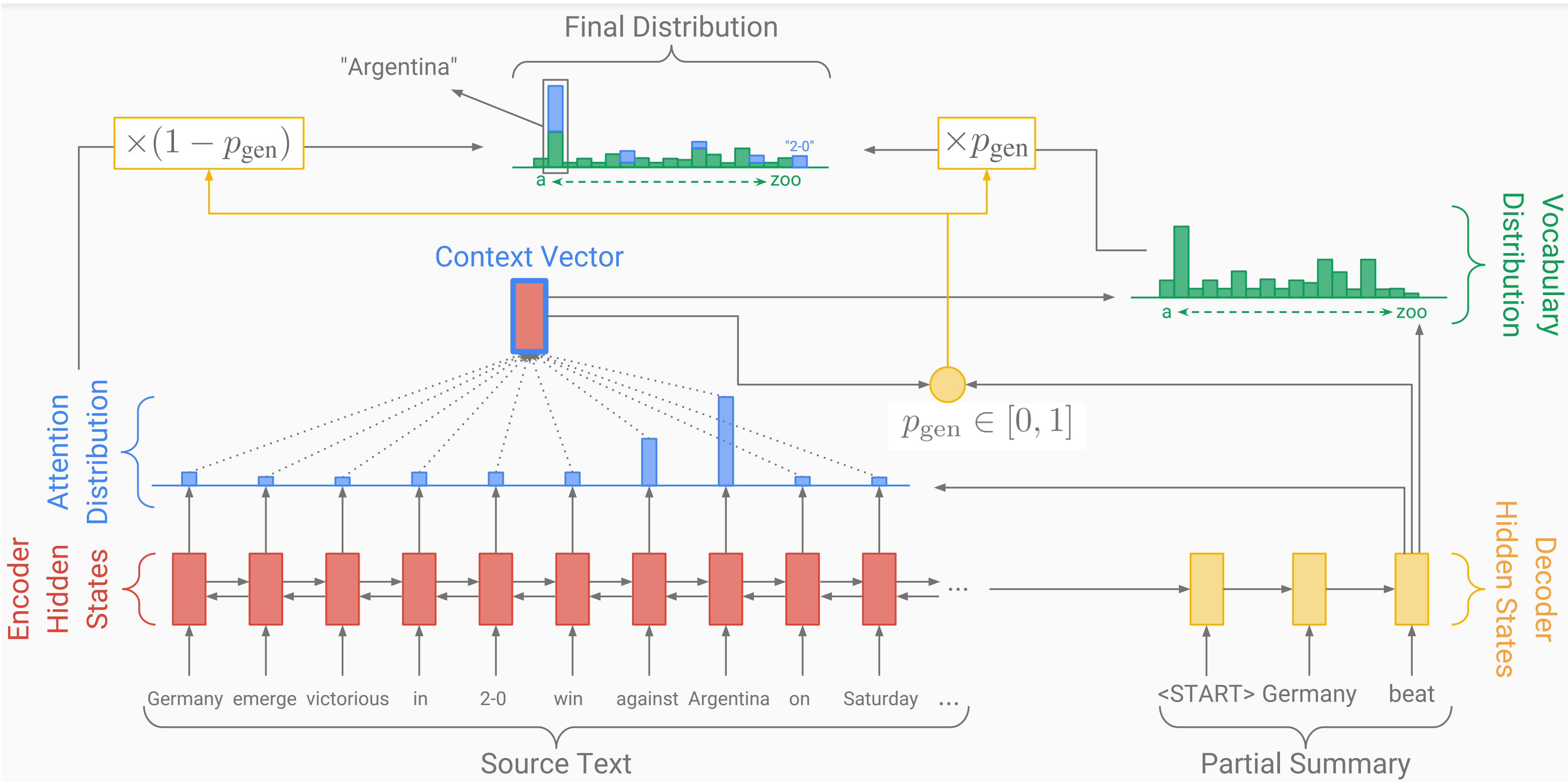
[Nallapati et al., 2016]

Generation+Copying



[Nallapati et al., 2016]

Pointer-Generator Networks



[See et al., 2017]

Coverage for Redundancy Reduction

Coverage = cumulative attention = what has been covered so far

Don't attend here

Source Text: Germany emerge victorious in 2-0 win against Argentina on Saturday

Summary: Germany beat _____

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]

Fast Reinforce-Selected Sentence Rewriting

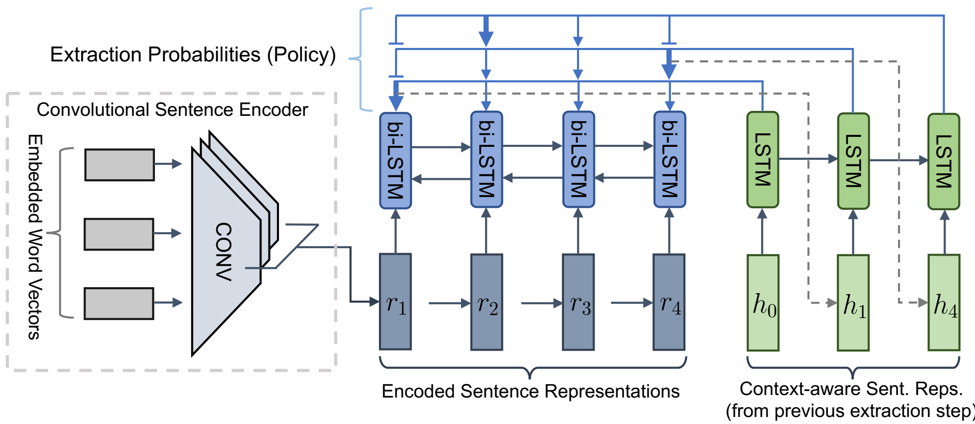


Figure 1: Our extractor agent: the convolutional encoder computes representation r_j for each sentence. The RNN encoder (blue) computes context-aware representation h_j and then the RNN decoder (green) selects sentence j_t at time step t. With j_t selected, h_{j_t} will be fed into the decoder at time $t+1$.

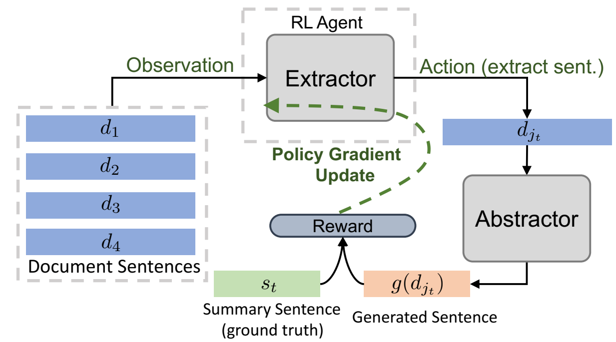


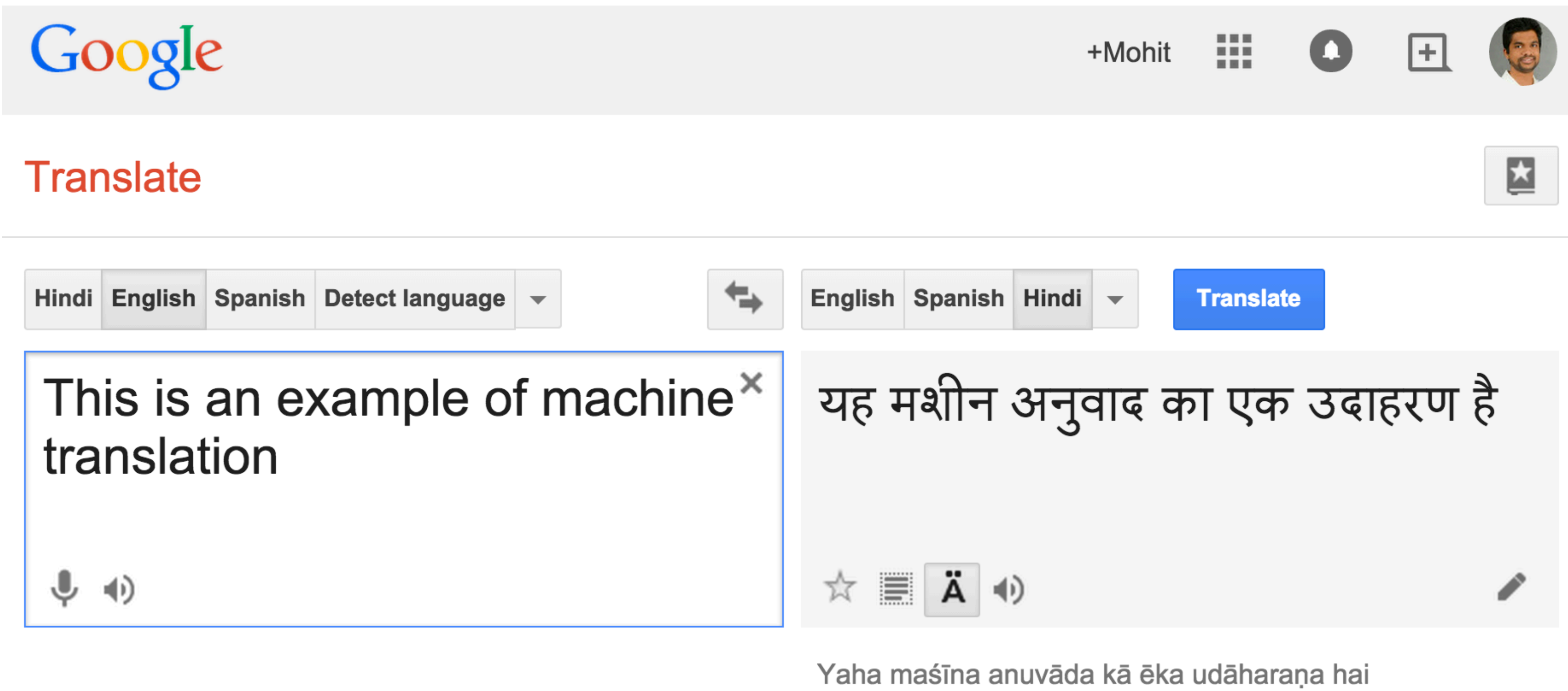
Figure 2: Reinforced training of the extractor (for one extraction step) and its interaction with the abstractor. For simplicity, the critic network is not shown. Note that all d's and s_t are raw sentences, *not* vector representations.

[Chen and Bansal, 2018]

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:

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

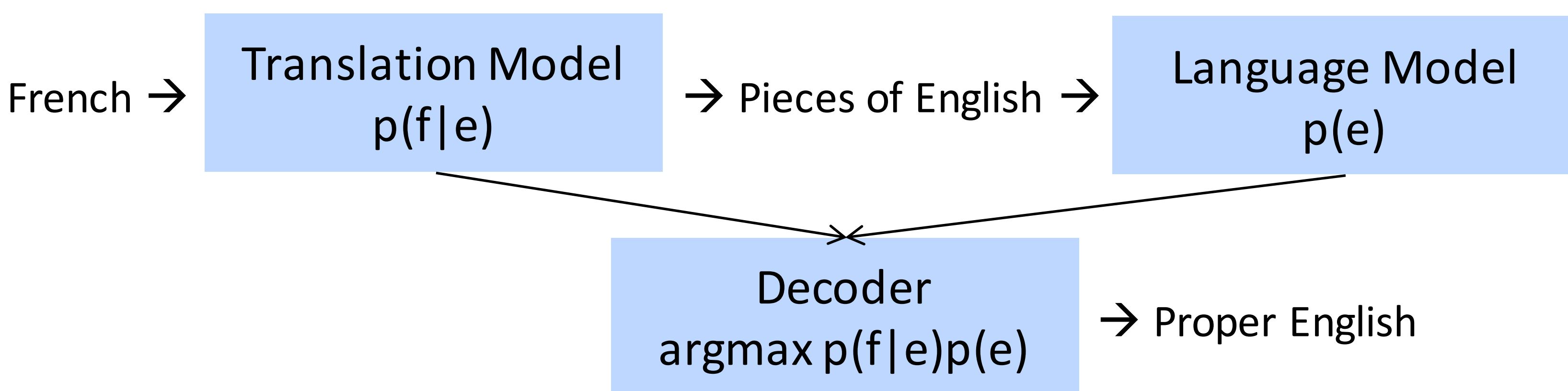
- Using Bayes rule, we get the following (since p(f) in the denominator is independent of the argmax over e):

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

[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!

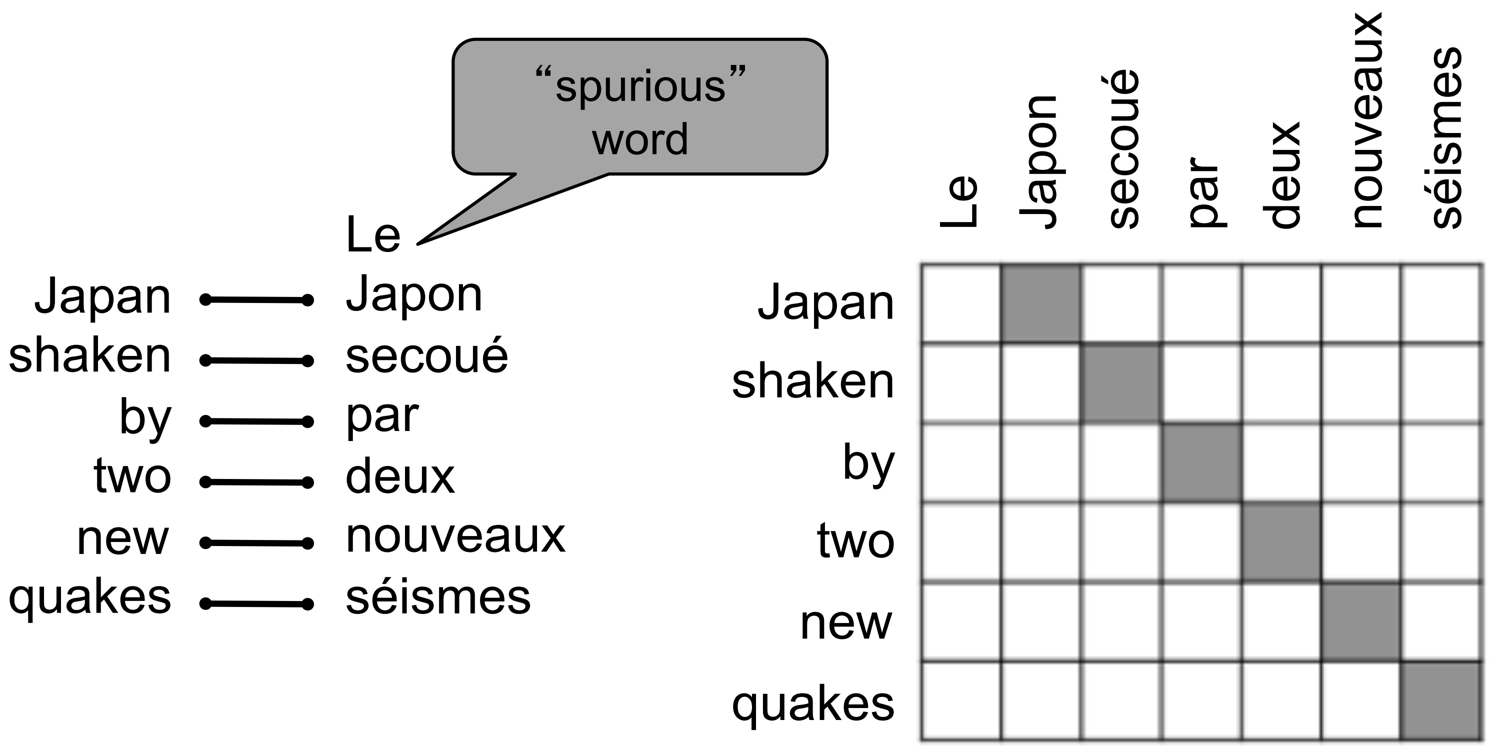


[Richard Socher CS224d]

Statistical Machine Translation

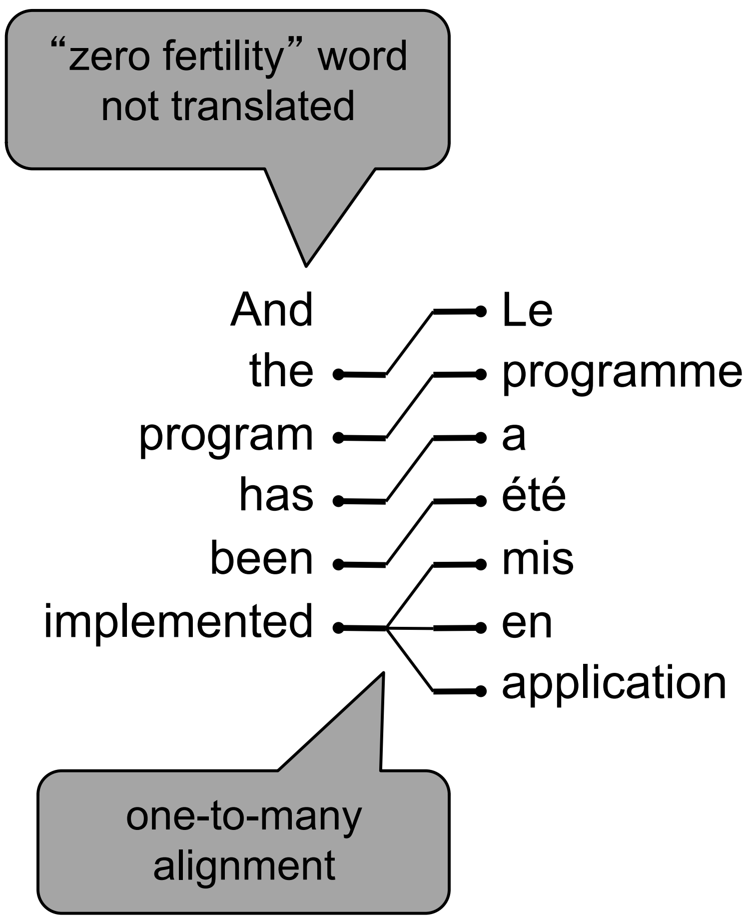
- First step in traditional machine translation is to find alignments or translational matchings between the two sentences, i.e., predict which words/phrases in French align to which words/phrases in English.

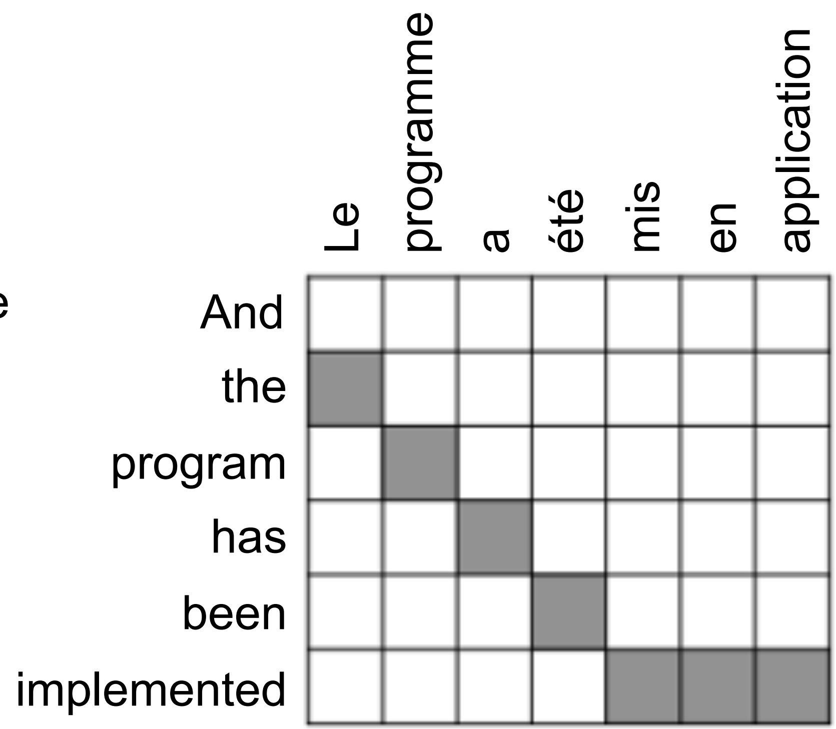
- Challenging problem: e.g., some words may not have any alignments:



Statistical Machine Translation

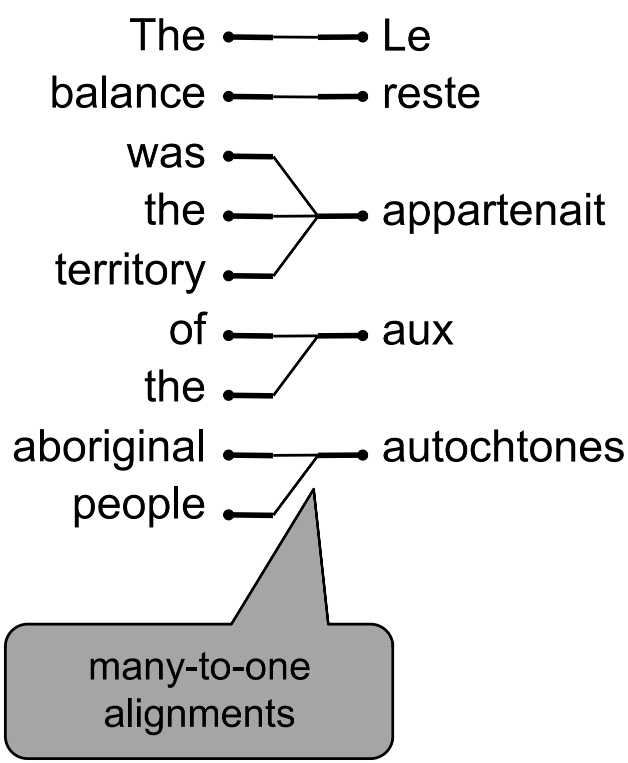
- One word in the source sentence might align to several words in the target sentence:

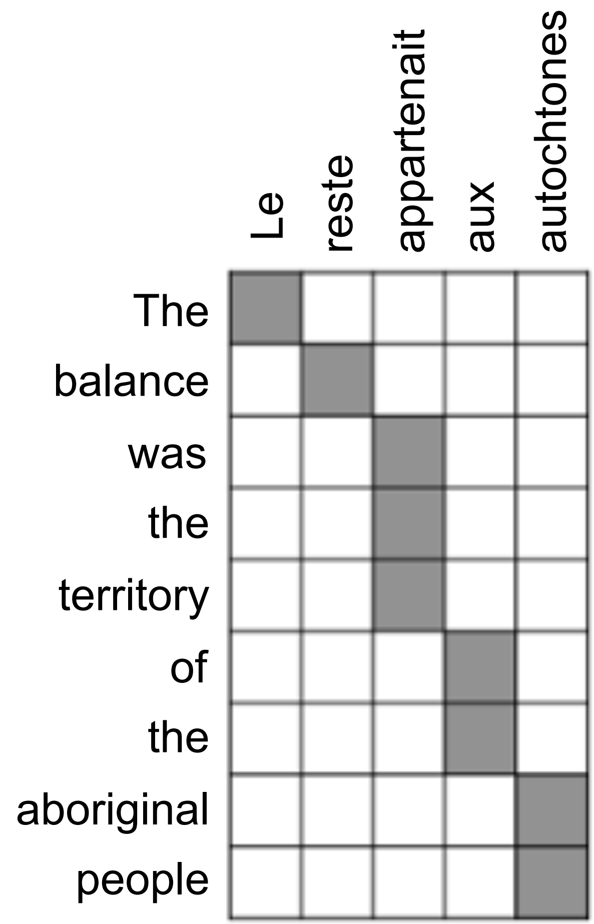




Statistical Machine Translation

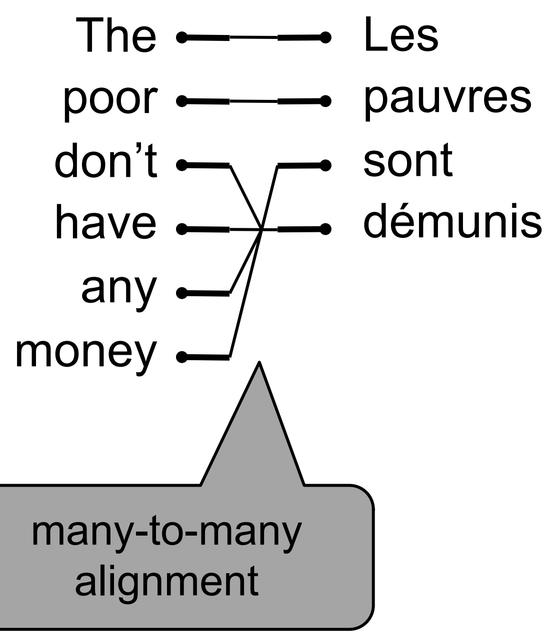
- Many words in the source sentence might align to a single word in the target sentence:

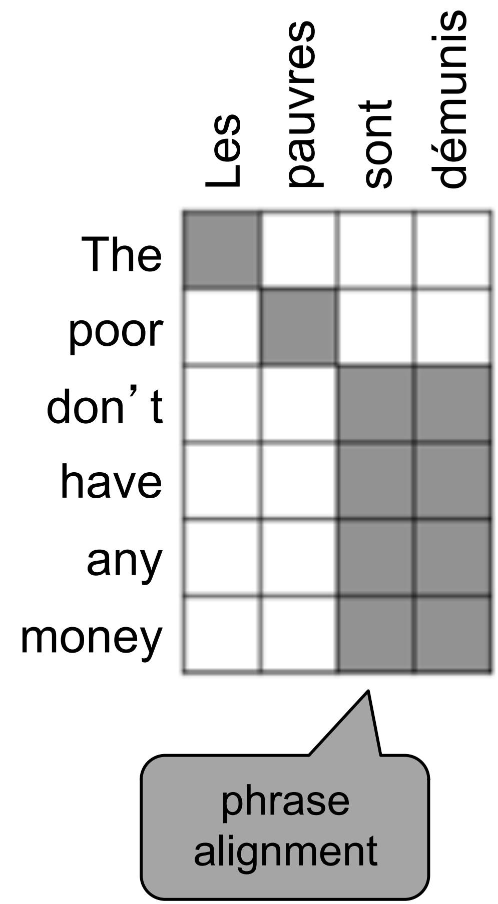




Statistical Machine Translation

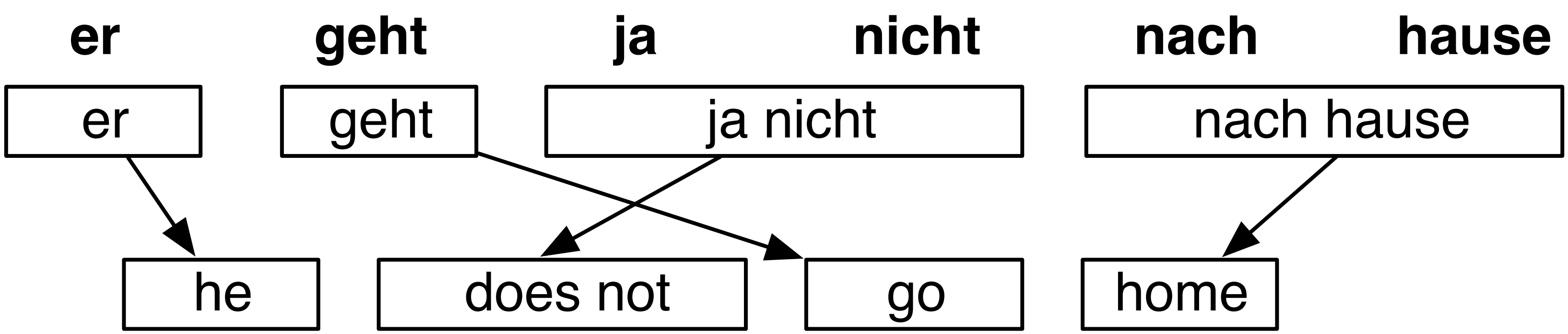
- And finally, many words in the source sentence might align to many words in the target sentence:





Statistical Machine Translation

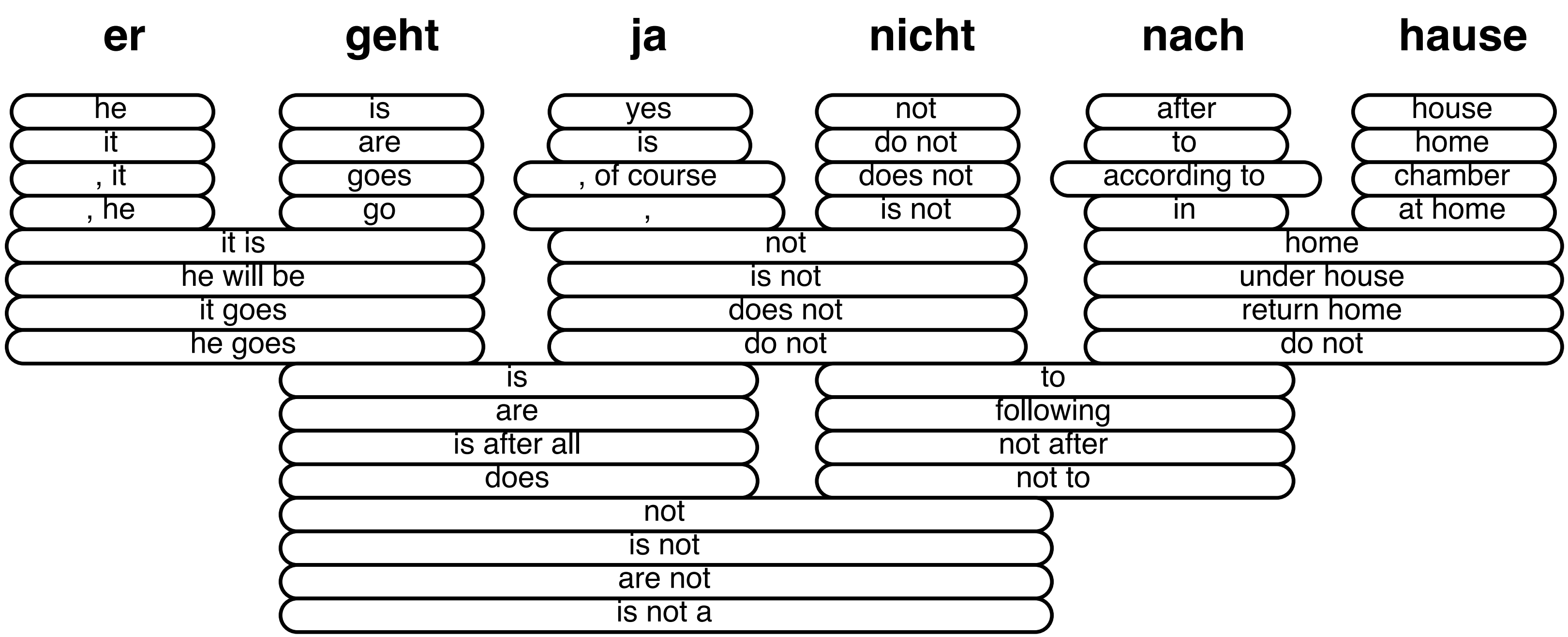
- After learning the word and phrase alignments, the model also needs to figure out the reordering, esp. important in language pairs with very different orders!



Statistical Machine Translation

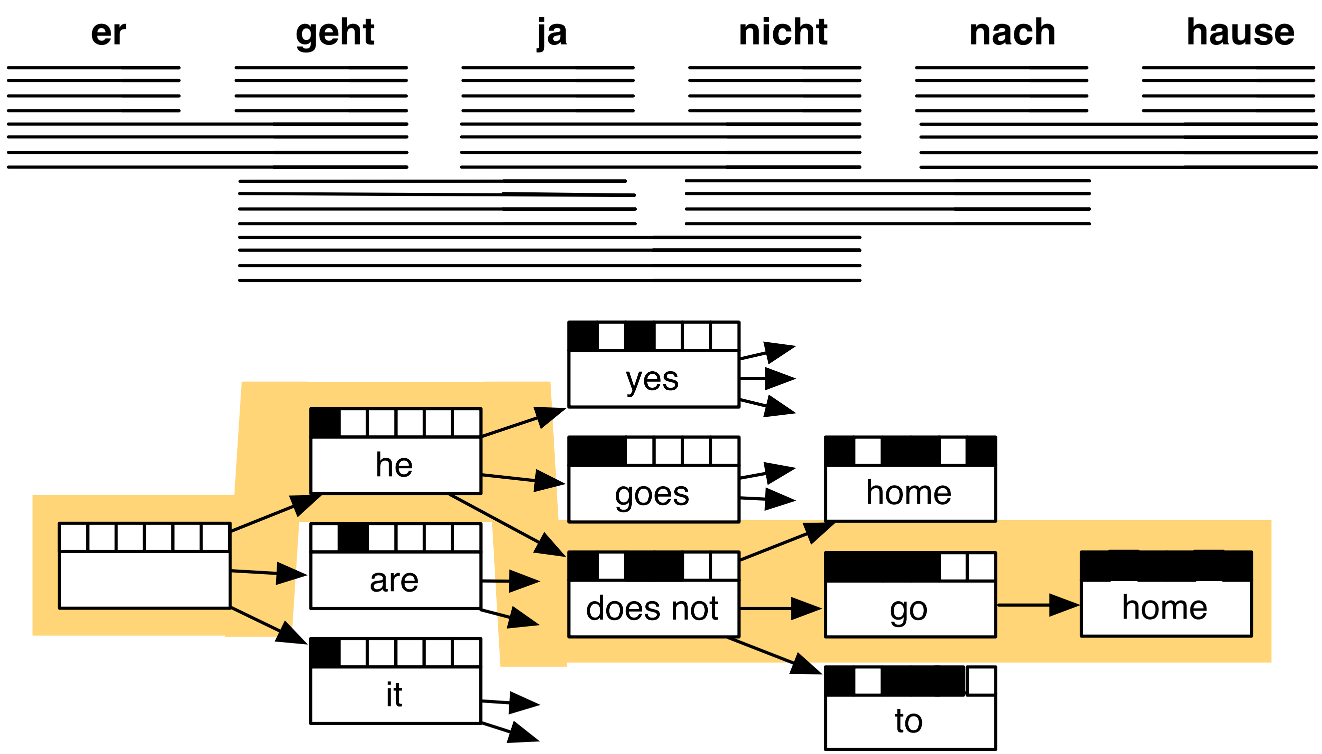
- After many steps, you get the large ‘phrase table’. Each phrase in the source language can have many possible translations in the target language, and hence the search space can be combinatorially large!

Translation Options



Statistical Machine Translation

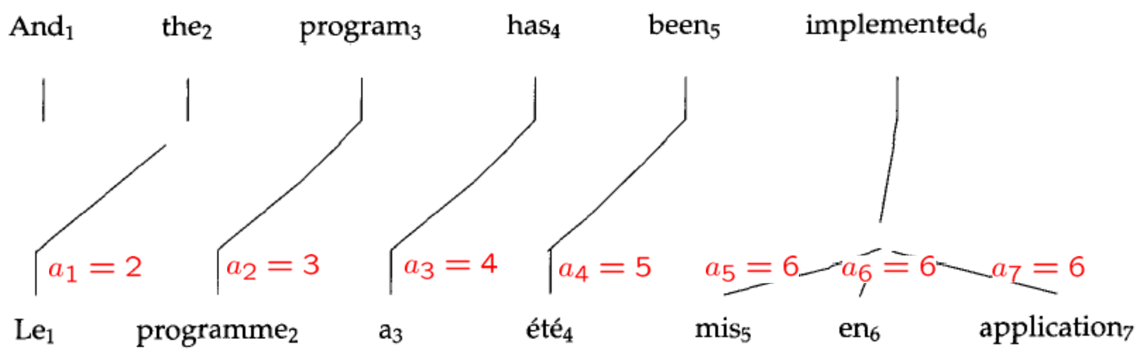
- Finally, you decode this hard search problem to find the best translation, e.g., using beam search on the several combinatorial paths through this phrase table (and also include the language model p(e) to rerank)



IBM Alignment Model 1

- Alignments: a hidden vector called an alignment specifies which English source is responsible for each French target word.
- The first, simplest IBM model treated alignment probabilities as roughly uniform:

$$a = a_1 \dots a_J$$



$$P(f, a|e) = \prod_j P(a_j = i) P(f_j|e_i)$$

$$= \prod_j \frac{1}{I+1} P(f_j|e_i)$$

$$P(f|e) = \sum_a P(f, a|e)$$

[Brown et al., 1993]

IBM Alignment Model 2 (Distortion)

- The next more advanced model captures the notion of ‘distortion’, i.e., how far from the diagonal is the alignment

$$P(f,a|e) = \prod_j P(a_j = i|j,I,J)P(f_j|e_i)$$

$$P(dist = i - j\frac{I}{J})$$

$$\frac{1}{Z}e^{-\alpha(i-j\frac{I}{J})}$$

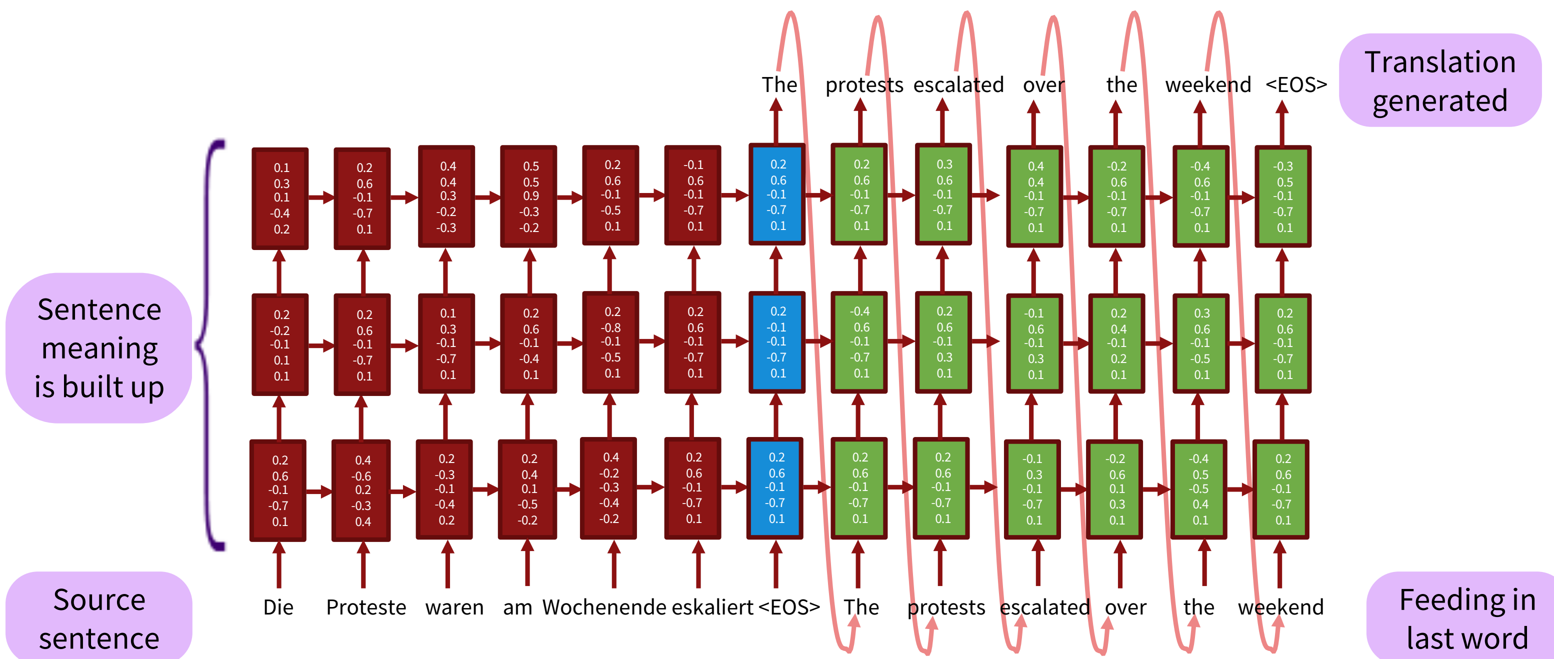
- Other approaches for biasing alignment towards diagonal include relative vs absolute alignment, asymmetric distances, and learning a full multinomial over distances
- Check the 4 other popular models (IBM3/4/5 and HMM model)

[Brown et al., 1993]

Neural Machine Translation

- Encoder-Decoder RNN models:

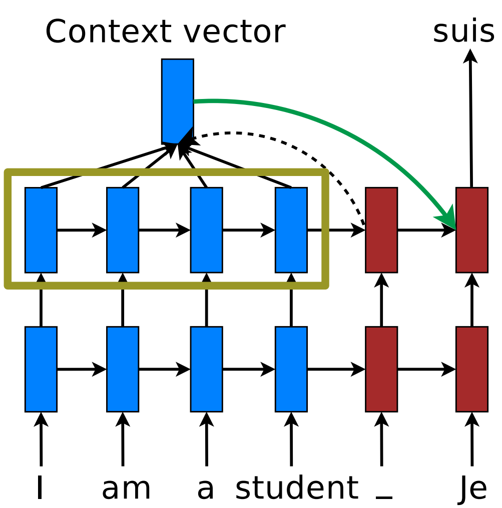
[Sutskever et al. 2014, Bahdanau et al. 2014, et seq.]
following [Jordan 1986] and more closely [Elman 1990]



A deep recurrent neural network

Alignment/Attention Models

- Translating longer sentences better, e.g., via attention/alignment module between encoder and decoder to jointly learn alignments and translations end-to-end

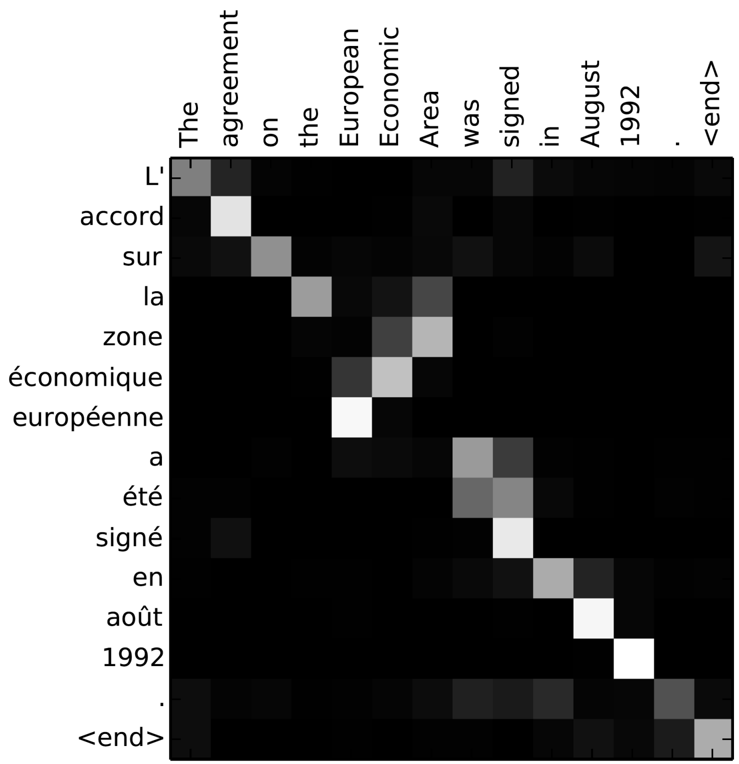


Bilinear form: well-adopted.

$$\mathrm{score}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_s) = \begin{cases} \boldsymbol{h}_t^\top \bar{\boldsymbol{h}}_s \\ \boldsymbol{h}_t^\top \boldsymbol{W_a} \bar{\boldsymbol{h}}_s \\ \boldsymbol{v}_a^\top \tanh\left(\boldsymbol{W_a}[\boldsymbol{h}_t; \bar{\boldsymbol{h}}_s]\right) \end{cases}$$

Alignment/Attention Models

- Translating longer sentences better, e.g., via attention/alignment module between encoder and decoder to jointly learn alignments and translations end-to-end

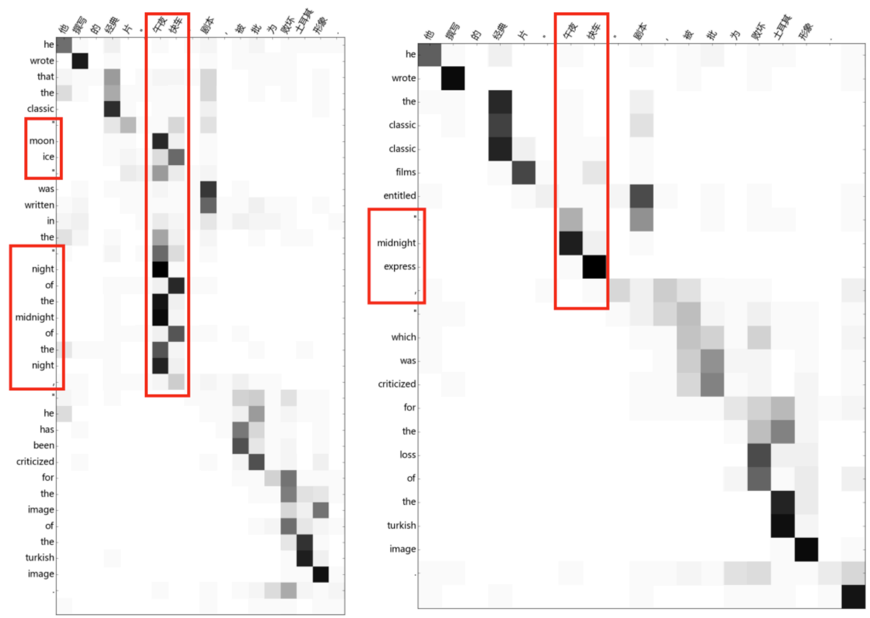


Dzmitry Bahdanau, KyungHuyn Cho, and Yoshua Bengio. ***Neural Machine Translation by Jointly Learning to Translate and Align****. ICLR'15.*

Linguistic Insights in NMT

Constraints on “distortion” (displacement) and fertility

➜ **Constraints on attention** [Cohn, Hoang, Vymolova, Yao, Dyer & Haffari NAACL 2016; Feng, Liu, Li, Zhou 2016 arXiv; Yang, Hu, Deng, Dyer, Smola 2016 arXiv].



Linguistic Insights in NMT

Extend to NMT – *Linguistic insights*

- [Cohn, Hoang, Vymolova, Yao, Dyer, Haffari, NAACL'16]: position (IBM2) + Markov (HMM) + fertility (IBM3-5) + alignment symmetry (BerkeleyAligner).

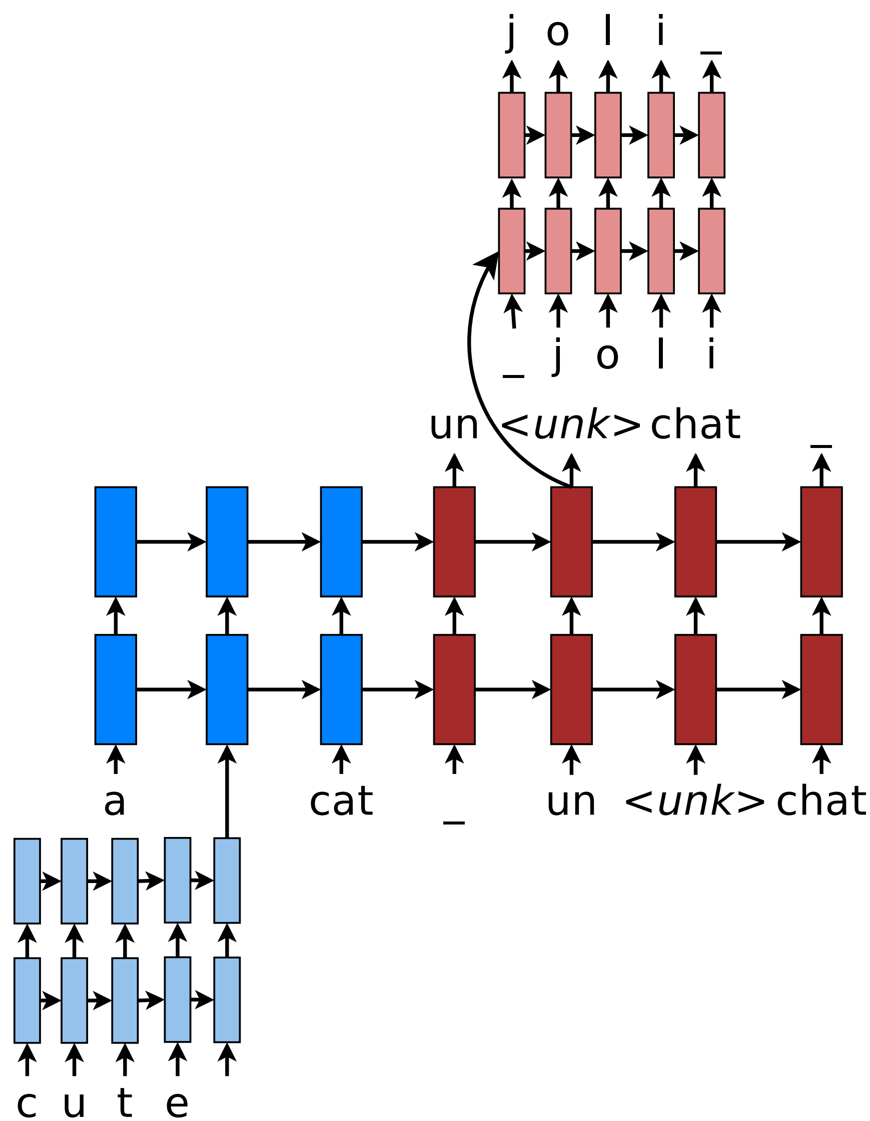
$$-\log(P(\mathbf{y}|\mathbf{x})) + \lambda \sum_{i}^{L} (1 - \sum_{t}^{C} \alpha_{ti})^2$$

Per source word

Source word fertility

- [Tu, Lu, Liu, Liu, Li, ACL'16]: linguistic & NN-based coverage models.

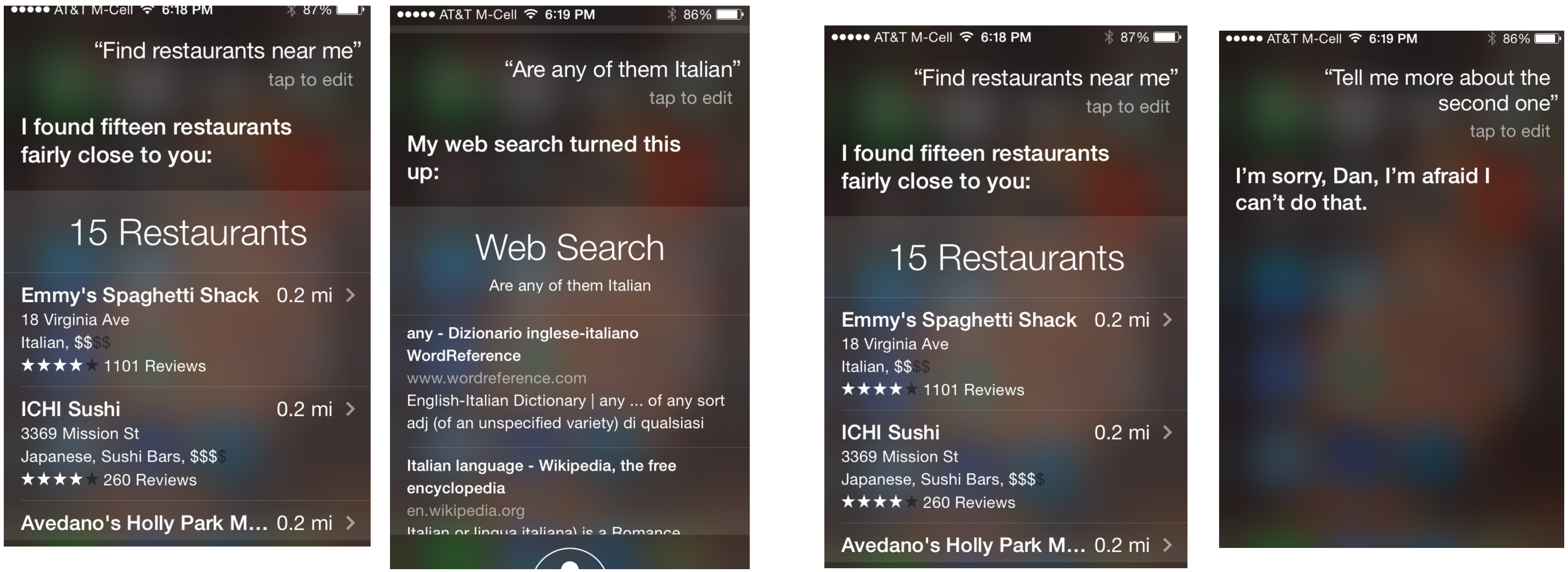
Hybrid Char-Word NMT



[Luong and Manning, 2016]

Dialogue Models

Examples

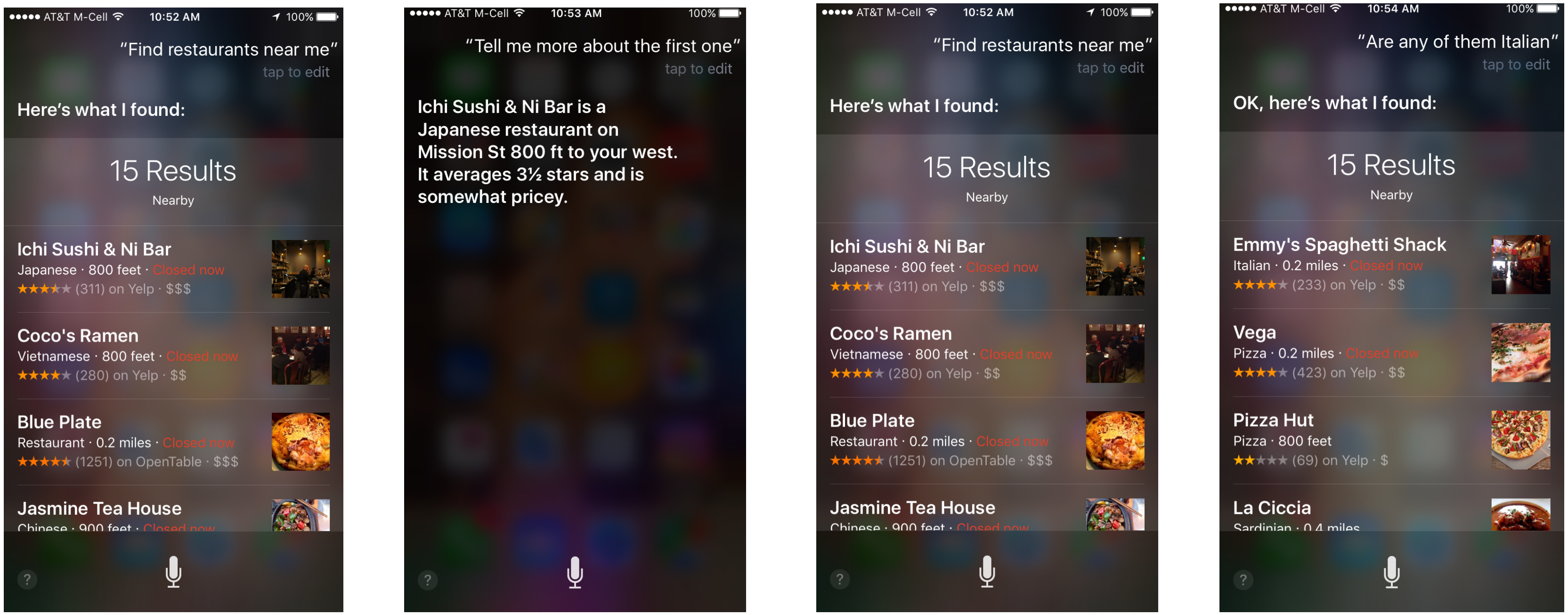


(a) (b)

Figure 29.1 Two sets of interactions with Siri in 2014. (a) A question ("Find restaurants near me") returns restaurants, but the system was unable to interpret a follow-up question ("Are any of them Italian?"). (b) An alternative followup ("Tell me more about the second one") similarly fails. This early system's confusion at follow-up questions suggests that it is mainly designed for a single interaction.

[Jurafsky SLP3 Book Chapter 29]

Examples



(a) (b)

Figure 29.2 The same follow-up questions that Siri couldn't answer in 2014 receive appropriate responses when posed to Siri in 2017.

Rule-based Chatbots (e.g., ELIZA)

```
Men are all alike.
IN WHAT WAY
They're always bugging us about something or other.
CAN YOU THINK OF A SPECIFIC EXAMPLE
Well  my boyfriend made me come here.
YOUR BOYFRIEND MADE YOU COME HERE
He says I'm depressed much of the time.
I AM SORRY TO HEAR YOU ARE DEPRESSED
..
..
WHO ELSE IN YOUR FAMILY TAKES CARE OF YOU?
My father
YOUR FATHER
You are like my father in some ways.
WHAT RESEMBLANCE DO YOU SEE
You are not very aggressive but I think you don't want me to notice that.
WHAT MAKES YOU THINK I AM NOT AGGRESSIVE
You don't argue with me.
WHY DO YOU THINK I DON'T ARGUE WITH YOU
You are afraid of me.
DOES IT PLEASE YOU TO BELIEVE I'M AFRAID OF YOU
My father is afraid of everybody.
WHAT ELSE COMES TO MIND WHEN YOU THINK OF YOUR  FATHER
Bullies.
DOES THAT HAVE ANYTHING TO DO WITH THE FACT THAT YOUR  BOYFRIEND MADE YOU COME HERE
```

Rule-based Chatbots (e.g., ELIZA)

```
(0 YOU 0 ME) [pattern]
->
(WHAT MAKES YOU THINK I 3 YOU) [transform]
```

function ELIZA GENERATOR(user *sentence*) **returns** *response*

Find the word *w* in *sentence* that has the highest keyword rank
  if *w* exists
    Choose the highest ranked rule *r* for *w* that matches *sentence*
    response ← Apply the transform in *r* to *sentence*
    if *w* = 'my'
      future ← Apply a transformation from the 'memory' rule list to *sentence*
      Push *future* onto memory stack
  else (no keyword applies)
   either
    response ← Apply the transform for the NONE keyword to *sentence*
   or
    response ← Pop the top response from the memory stack
  return(*response*)

Figure 29.5 A simplified sketch of the ELIZA algorithm. The power of the algorithm come from the particular transforms associated with each keyword.

IR-based Chatbots

- Retrieval systems use two major approaches to “extract” the best response from a dialogue corpus, given the new, test-time user utterance:

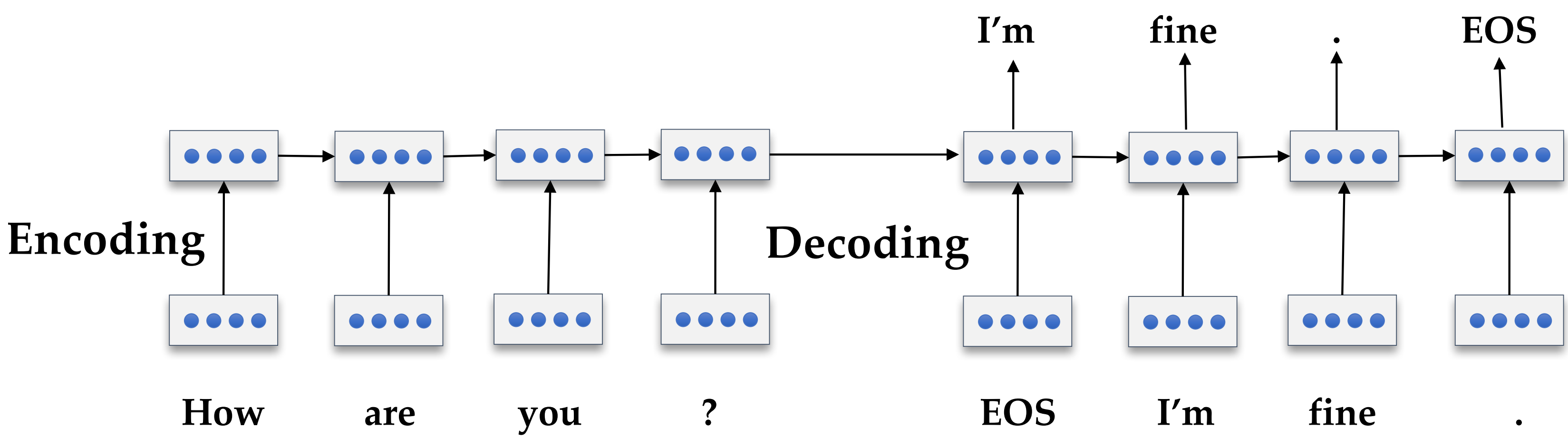
- **1) Return Response of Most Similar Turn:** Find conversation turn *t* (in corpus *C*) which is most similar to the given user utterance/query *q*, and return the following turn/response *r* of that most-similar utterance:

$$r = response\left(\underset{t \in C}{\operatorname{argmax}} \frac{q^T t}{||q||t||}\right)$$

- **2) Return Most Similar Turn:** Instead of returning the following turn of the most similar utterance, we return this most similar utterance itself, with the intuition that a good response often shared words/semantics with the prior turn:

$$r = \underset{t \in C}{\operatorname{argmax}} \frac{q^T t}{||q||t||}$$

Seq-to-Seq Chatbots



[Shang et al. 2015; Vinyals and Le, 2015; Sordoni et al., 2015]

Evaluating Chatbots

- Automatic metrics based on word/phrase overlap not very useful because so many responses might be correct/appropriate for chitchat (task-based dialogue easier/more meaningful to evaluate via success/completion rate, etc.)

- Human evaluation most meaningful/common (but time-consuming + expensive)

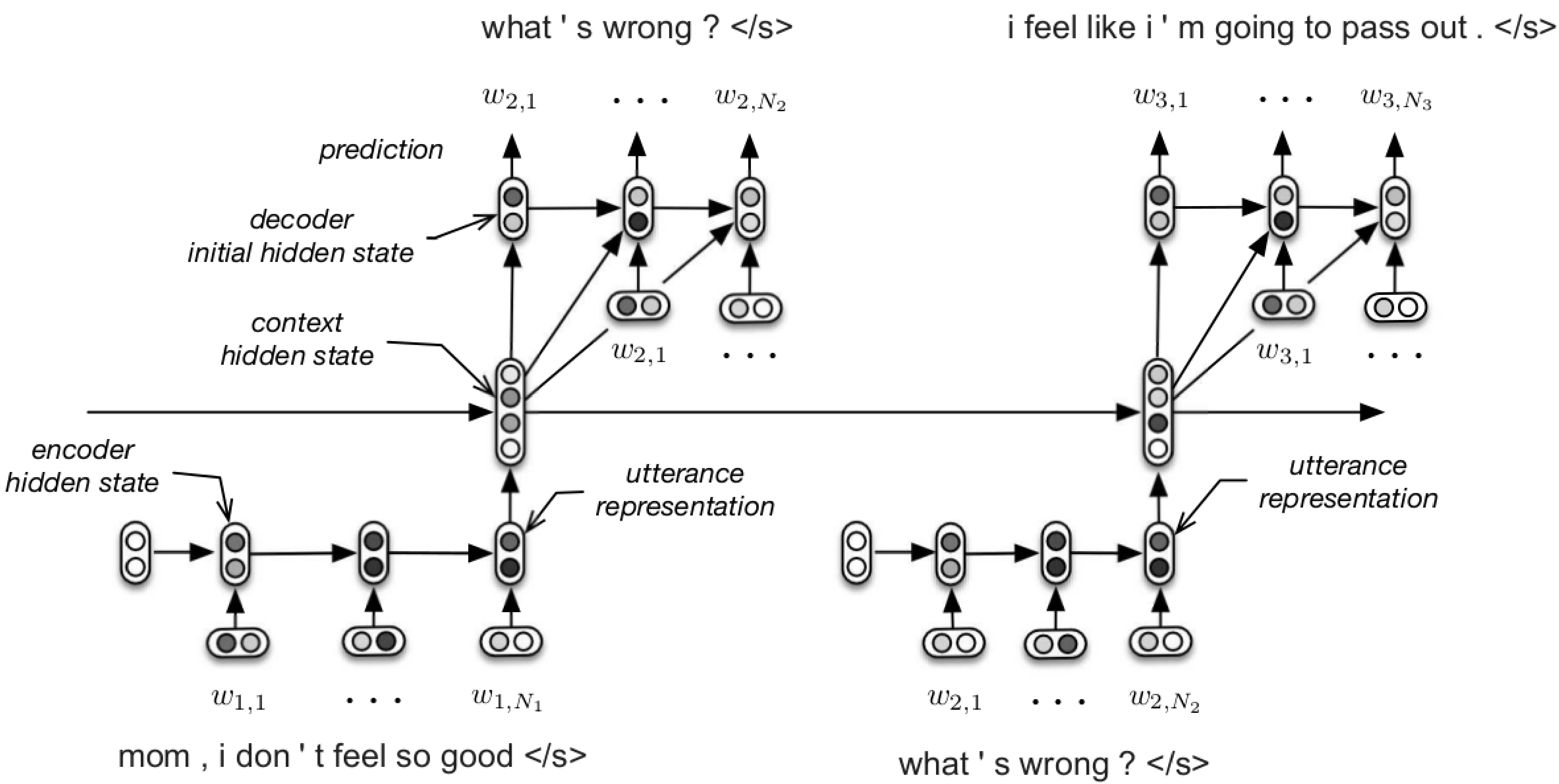
- Can't do slot-filling techniques because this is not task-oriented dialogue with a specific goal or success metric

- Engagement or length of conversation in real human-based setup?

- Some new automatic classification approaches like ADEM [Lowe et al., 2017] to classify appropriateness of response, and Adversarial evaluation [Bowman et al., 2016; Kannan and Vinyals, 2016; Li et al., 2017] to fool a classifier that distinguishes between human and machine generated responses

Some Advanced Seq-to-Seq Models

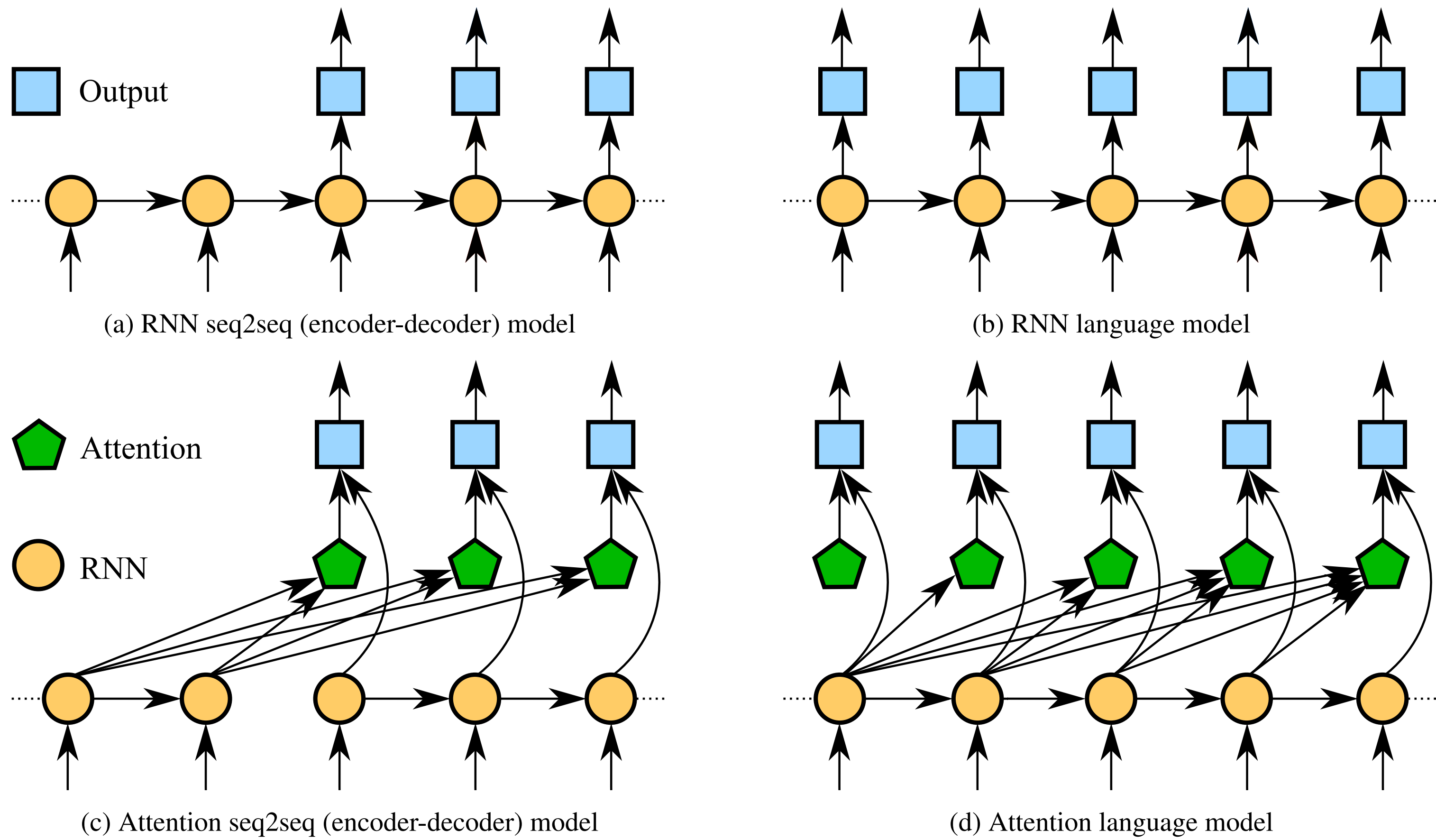
- Hierarchical Recurrent Encoder-Decoder



[Serban et al., 2015]

Some Advanced Seq-to-Seq Models

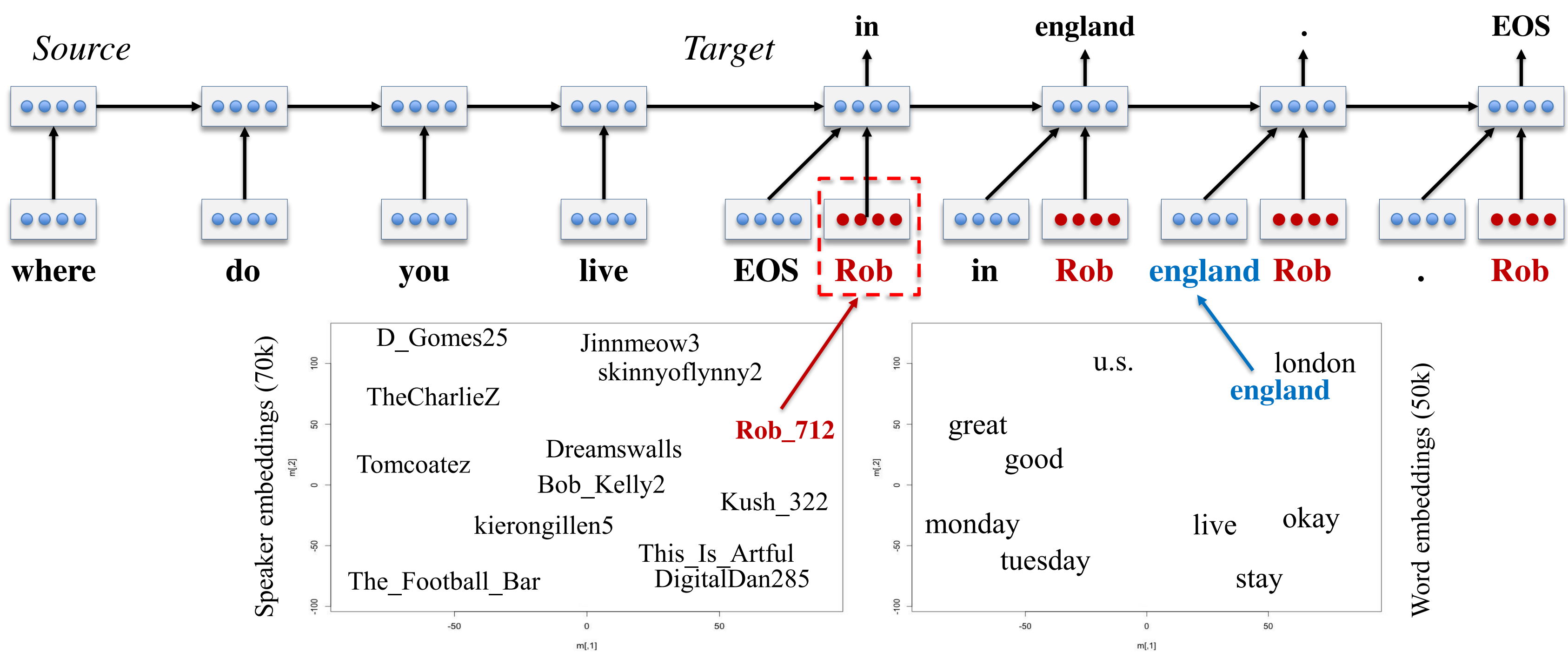
- Attention-RNN Language Model



[Mei et al., 2016]

Some Advanced Seq-to-Seq Models

- Persona-based Language Models



[Li et al., 2016]

Polite/Rude Reinforced Dialogue Models

- Rewards for making model more polite

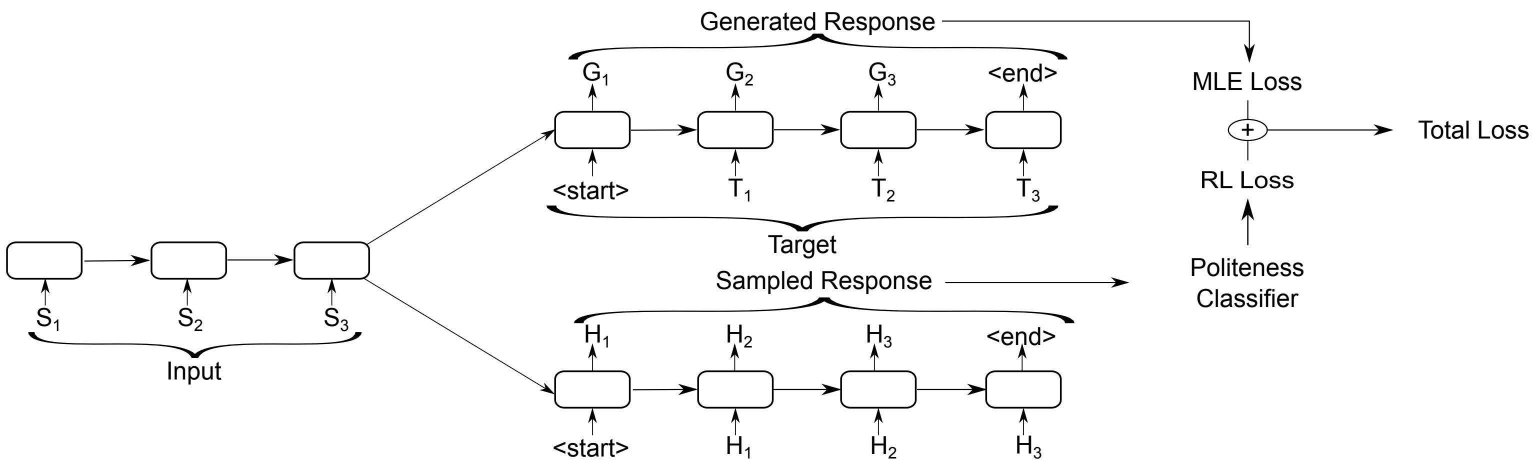


Figure 4: Polite-RL model: upper-right shows max-likelihood (ML) training with generated and ground-truth target sequences; lower-right shows RL training with a randomly sampled response generated by the model and the reward it generates after getting fed into the style classifier. Note that the attention mechanism is not shown here for clarity.

[Niu and Bansal, 2018]

Frame-based & KB-based Dialog Agents

GUS:	Hello. My name is GUS. I can help you plan a simple trip by air. Where do you want to go ?
Client:	I want to go to San Diego on May 28
GUS:	What time do you want to leave?
Client:	I must be in San Diego before 10 am
GUS:	Would you like P.S.A. flight 102 that arrives at 9.15 am ?
Client:	What is the next flight ?
GUS:	Air California flight 310 that leaves at 8.30am and arrives at 10.00am
Client:	I'll take the first one
GUS:	In what name should I make a reservation, please?
Client:	Dan Bobrow
GUS:	I have confirmed the following flight:
Client:	P.S.A. flight 102 on Wednesday May 28 from San Jose to San Diego that leaves at 7.30 am and arrives at 9.15 am.
GUS:	What date do you want to return on ?
Client:	On Friday in the evening.
GUS:	Would you like the flight that leaves at 7.45 pm ?
Client:	That's fine.
GUS:	I have confirmed the following flight: P.S.A. flight 307 on Friday May 30 from San Diego to San Jose that leaves at 7.45 pm and arrives at 9.30 pm Thank you for calling. Goodbye

Figure 29.8 The travel domain: A transcript of an actual dialog with the GUS system of Bobrow et al. (1977). P.S.A. and Air California were airlines of that period.

[Jurafsky SLP3]

Frame-based & KB-based Dialog Agents

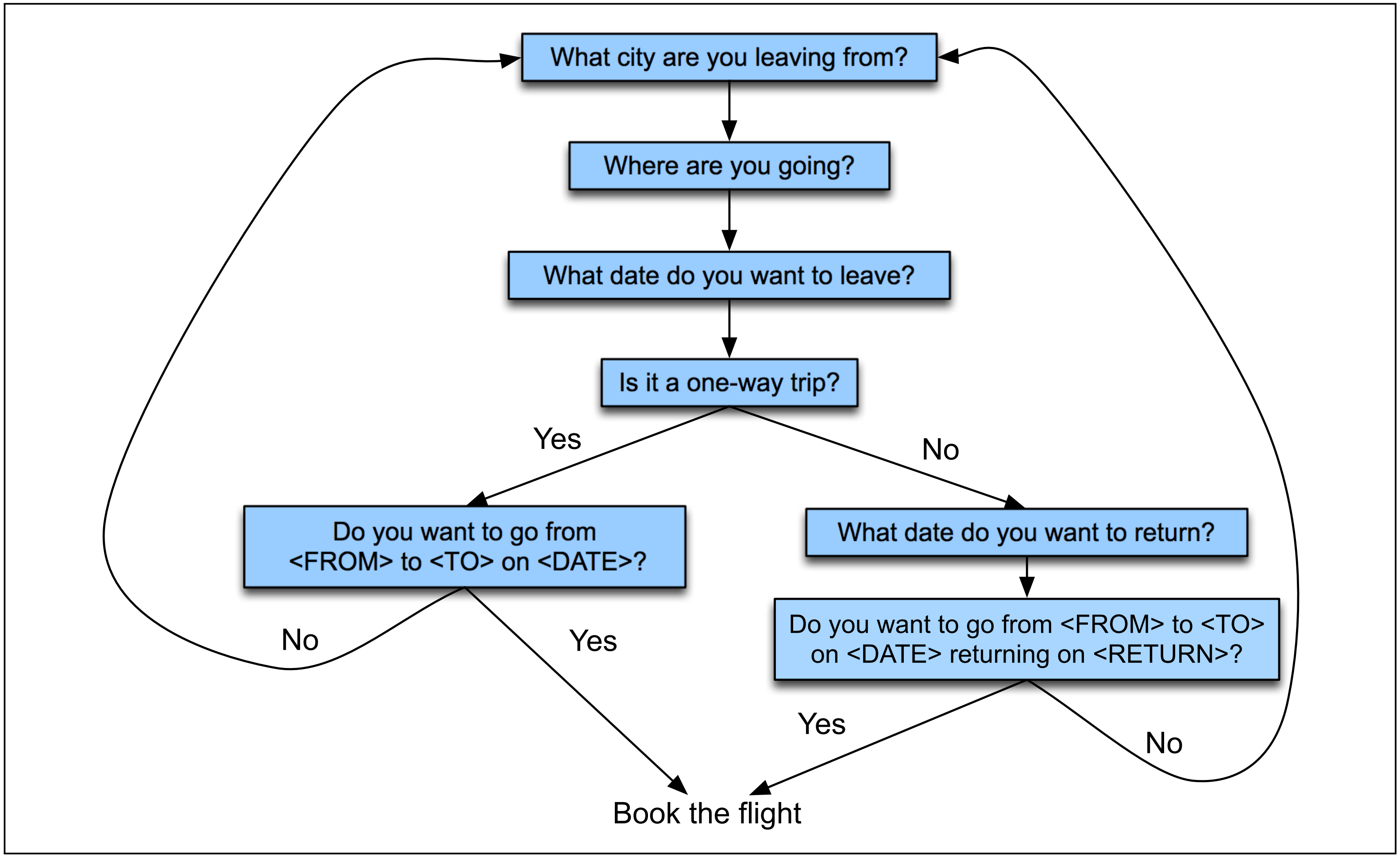


Figure 29.9 A simple finite-state automaton architecture for frame-based dialog.

Frame-based & KB-based Dialog Agents

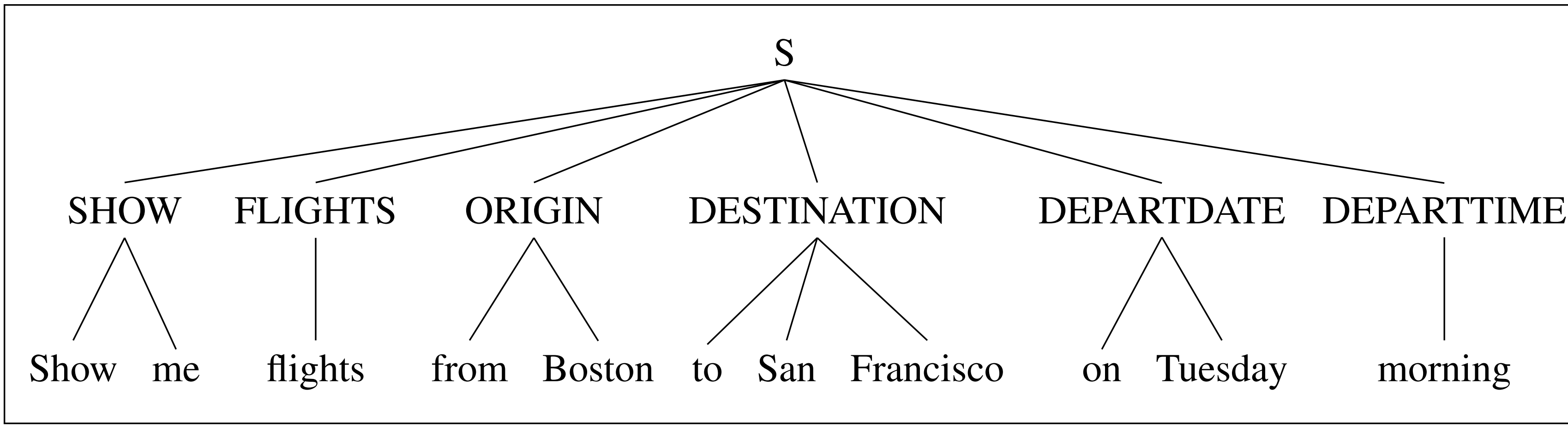


Figure 29.10 A semantic grammar parse for a user sentence, using slot names as the internal parse tree nodes.

Frame-based & KB-based Dialog Agents

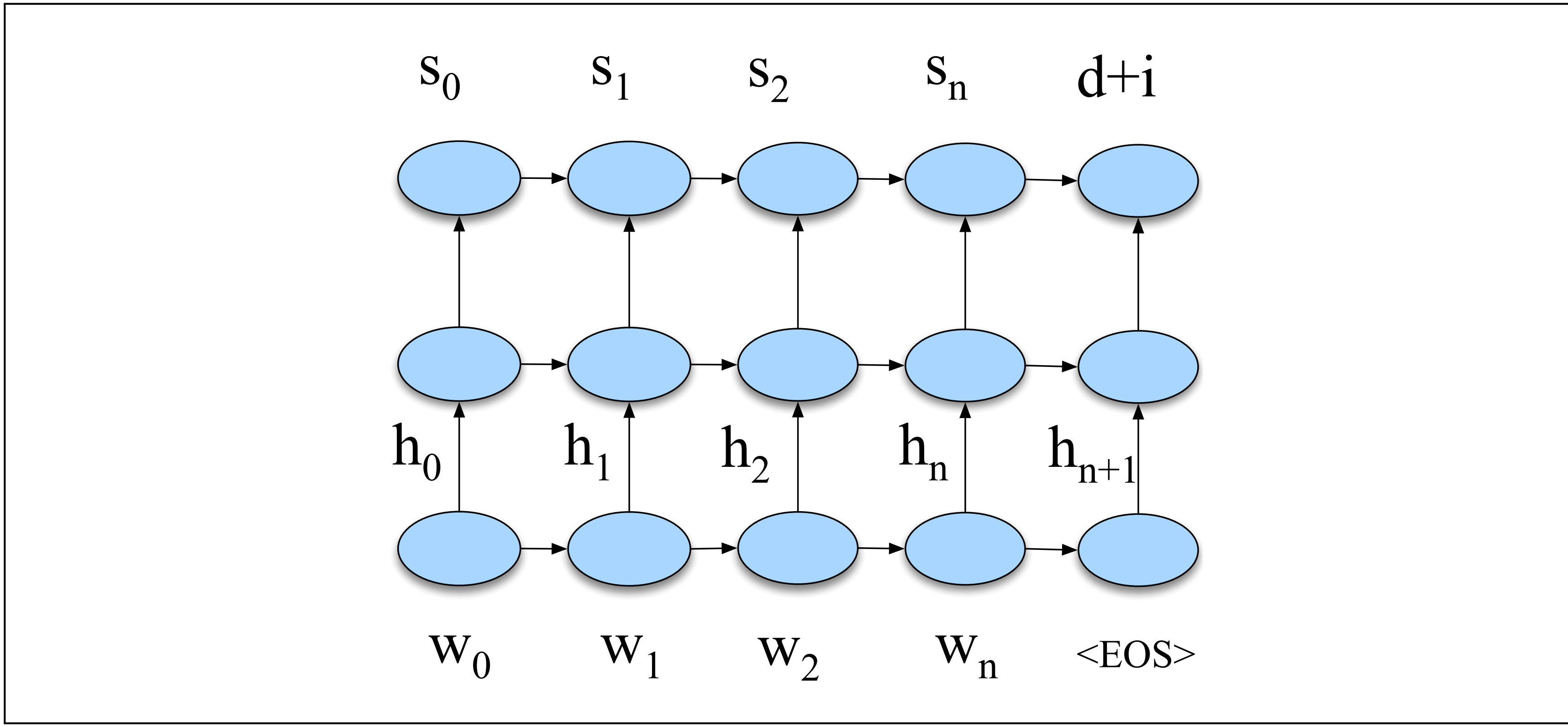


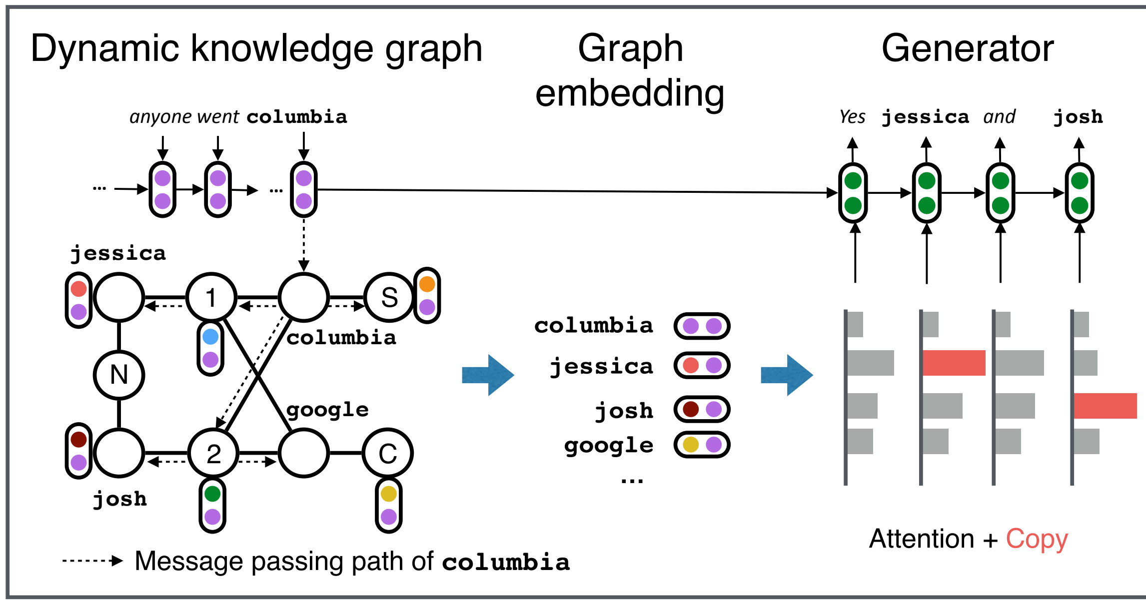
Figure 29.11 An LSTM architecture for slot filling, mapping the words in the input (represented as 1-hot vectors or as embeddings) to a series of IOB tags plus a final state consisting of a domain concatenated with an intent.

Frame-based & KB-based Dialog Agents

KB + Dialogue history

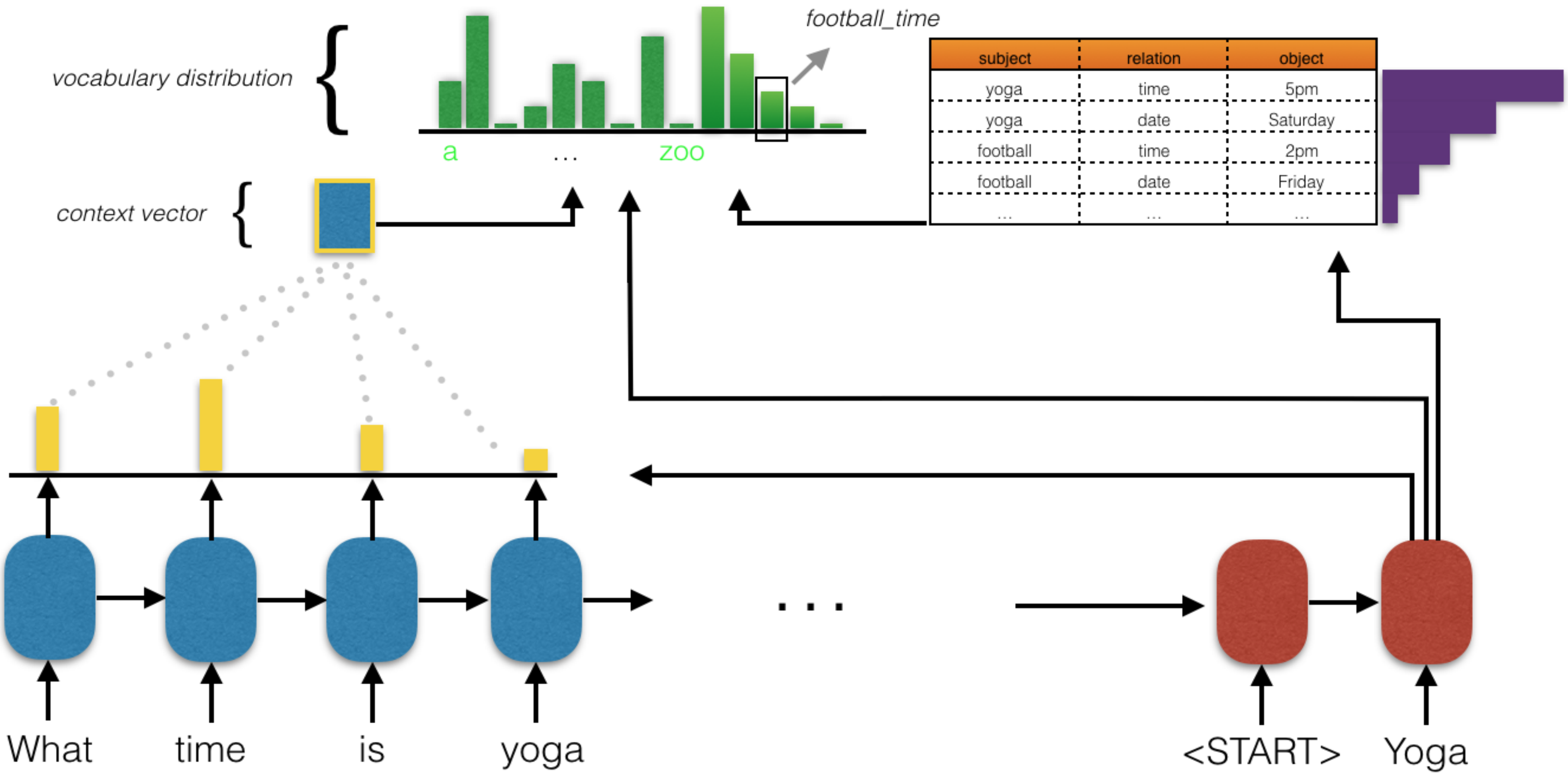
	Name	School	Company
Item 1	Jessica	Columbia	Google
Item 2	Josh	Columbia	Google

B: *anyone went to* **columbia**?



[He et al., 2017]

Frame-based & KB-based Dialog Agents



[Eric et al., 2017]