COMP 786(Fall 2020)
Natural Language Processing

Week 7: Semantic Parsing 2; Question Answering

THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL

Mohit Bansal

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

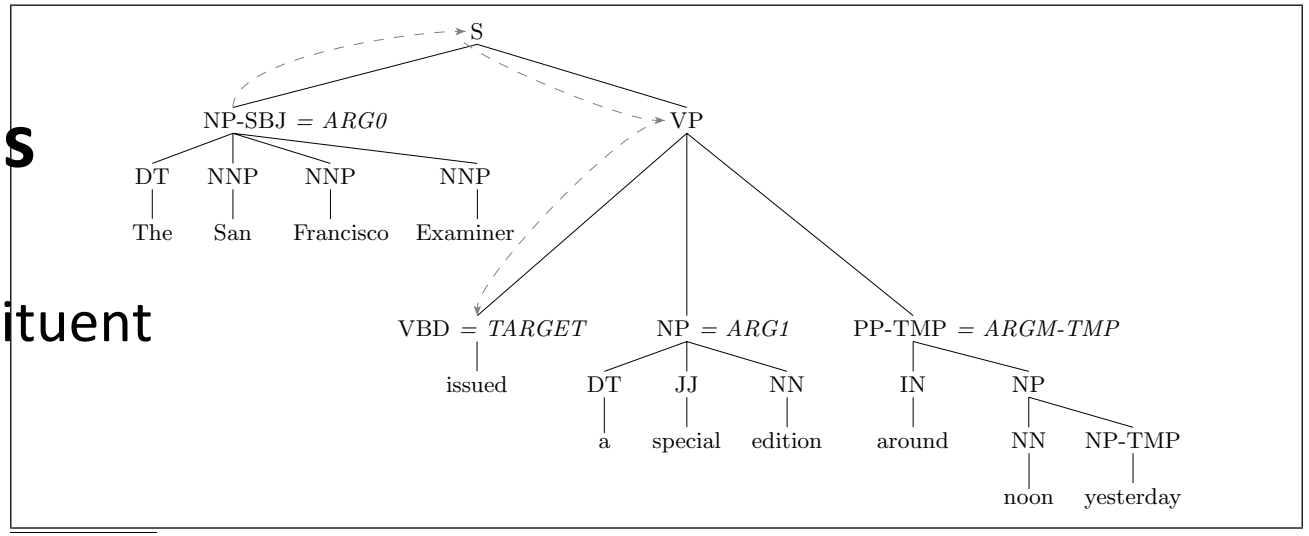
Announcements

- Coding-HW1 (on word vector training+evaluation_ +visualization) due today midnight!
- Midterm project presentation next week (look for details in my email last week).
- Please send 1-2 paragraph project description asap today if you haven't yet (was due yesterday Sep21).

SRL and Semantic Parsing 2
(AMR, Neural Models, etc.)

SRL Features

Features



Headword of constituent

Examiner

Headword POS

NNP

Voice of the clause

Active

Subcategorization of pred

VP -> VBD NP PP

Named Entity type of constit

ORGANIZATION

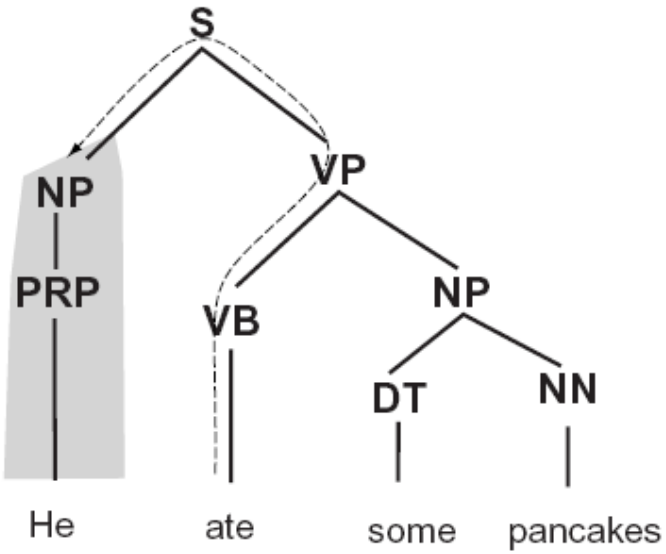
First and last words of constit

The, Examiner

Linear position,clause re: predicate

before

Path-based Features for SRL



Path	*Description*
VB↑VP↓PP	PP argument/adjunct
VB↑VP↑S↓NP	subject
VB↑VP↓NP	object
VB↑VP↑VP↑S↓NP	subject (embedded VP)
VB↑VP↓ADVP	adverbial adjunct
NN↑NP↑NP↓PP	prepositional complement of noun

Some SRL Results

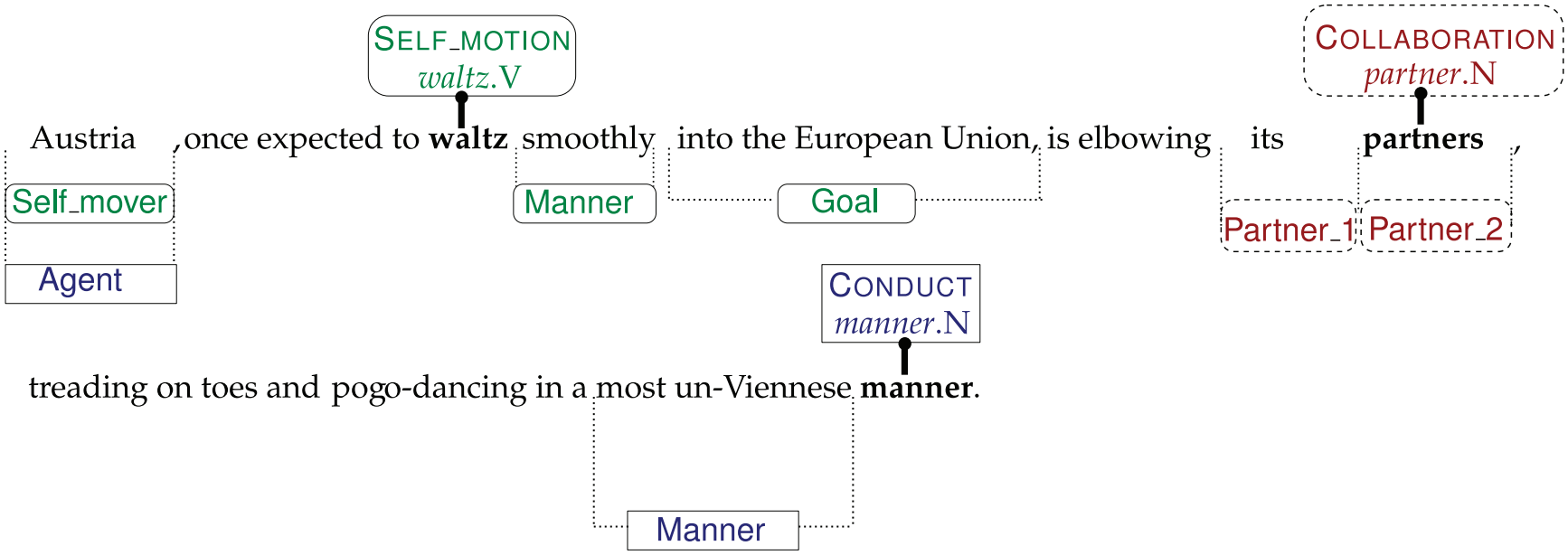
- So major feature categories in traditional feature-based SRL models were:
  - Headword, syntactic type, case, etc. of candidate node/ constituent
  - Linear and tree path from predicate target to node
  - Active vs. passive voice
  - Second order and higher order features

- Accuracy for such feature-based SRL models then highly depends on accuracy of underlying parse tree!
  - So quite high SRL results when using ground-truth parses
  - Much lower results with automatically-predicted parses!

CORE		ARGM	
F1	Acc.	F1	Acc.
92.2	80.7	89.9	71.8

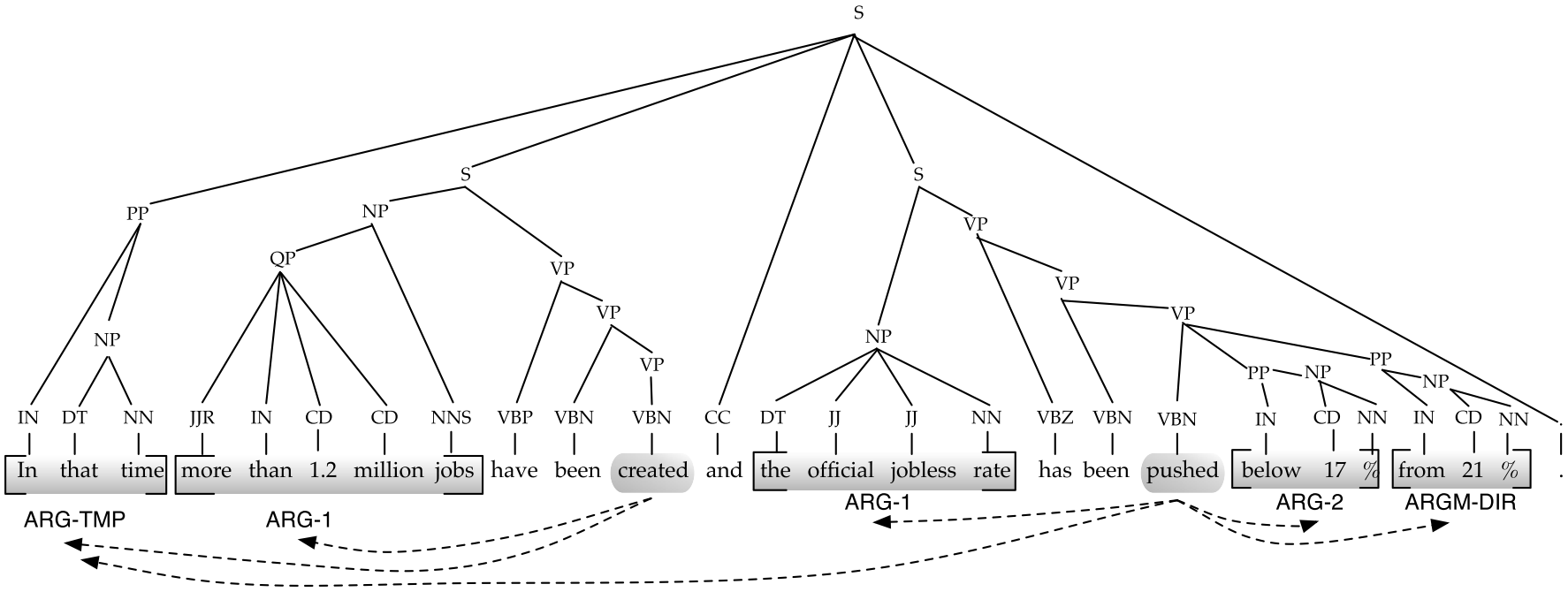
CORE		ARGM	
F1	Acc.	F1	Acc.
84.1	66.5	81.4	55.6

Schematic of Frame Semantics (FrameNet)

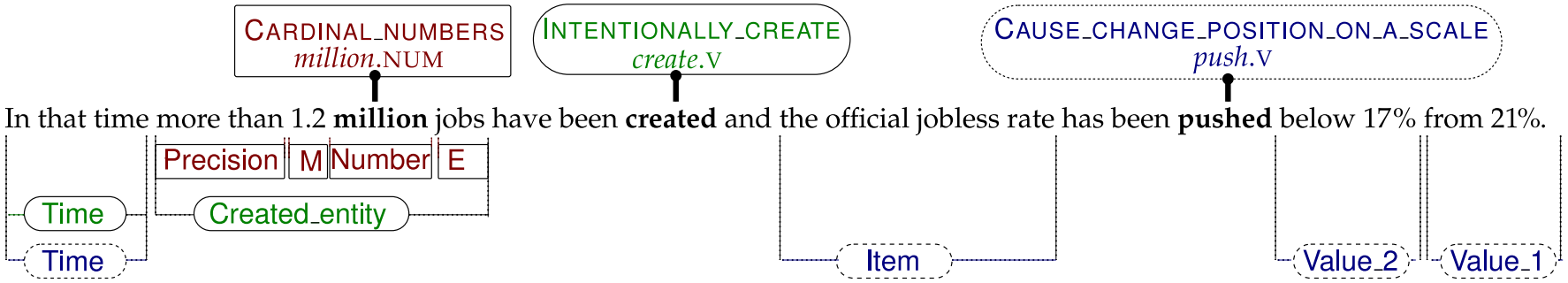


[Das et al., 2014]

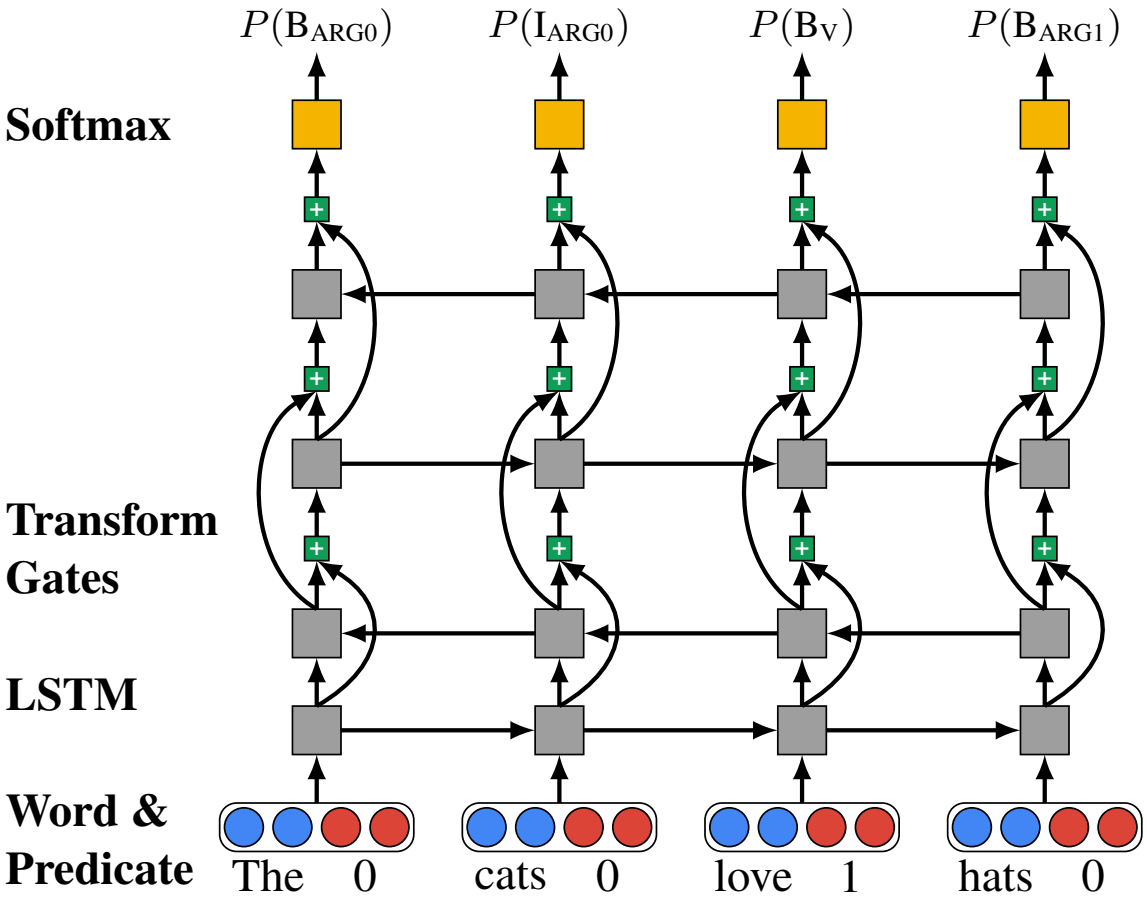
PropBank vs. FrameNet Representations



(a)



Neural SRL



[He et al., 2017]

Neural SRL

- CoNLL 2005 dataset:

	Development				WSJ Test				Brown Test				Combined
Method	P	R	F1	Comp.	P	R	F1	Comp.	P	R	F1	Comp.	F1
Ours (PoE)	**83.1**	**82.4**	**82.7**	**64.1**	**85.0**	**84.3**	**84.6**	**66.5**	**74.9**	**72.4**	**73.6**	**46.5**	**83.2**
Ours	81.6	81.6	81.6	62.3	83.1	83.0	83.1	64.3	72.9	71.4	72.1	44.8	81.6
Zhou	79.7	79.4	79.6	-	82.9	82.8	82.8	-	70.7	68.2	69.4	-	81.1
FitzGerald (Struct.,PoE)	81.2	76.7	78.9	55.1	82.5	78.2	80.3	57.3	74.5	70.0	72.2	41.3	-
Täckström (Struct.)	81.2	76.2	78.6	54.4	82.3	77.6	79.9	56.0	74.3	68.6	71.3	39.8	-
Toutanova (Ensemble)	-	-	78.6	58.7	81.9	78.8	80.3	60.1	-	-	68.8	40.8	-
Punyakanok (Ensemble)	80.1	74.8	77.4	50.7	82.3	76.8	79.4	53.8	73.4	62.9	67.8	32.3	77.9

- CoNLL 2012 dataset:

	Development				Test			
Method	P	R	F1	Comp.	P	R	F1	Comp.
Ours (PoE)	**83.5**	**83.2**	**83.4**	**67.5**	**83.5**	**83.3**	**83.4**	**68.5**
Ours	81.8	81.4	81.5	64.6	81.7	81.6	81.7	66.0
Zhou	-	-	81.1	-	-	-	81.3	-
FitzGerald (Struct.,PoE)	81.0	78.5	79.7	60.9	81.2	79.0	80.1	62.6
Täckström (Struct.)	80.5	77.8	79.1	60.1	80.6	78.2	79.4	61.8
Pradhan (revised)	-	-	-	-	78.5	76.6	77.5	55.8

[He et al., 2017]

Neural SRL

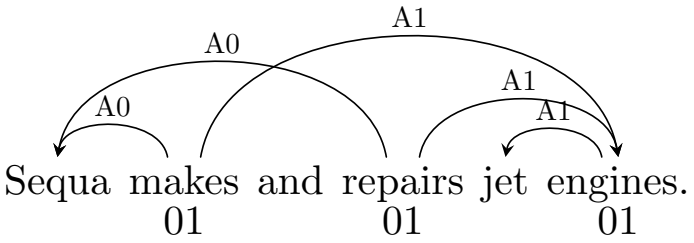


Figure 1: A semantic dependency graph.

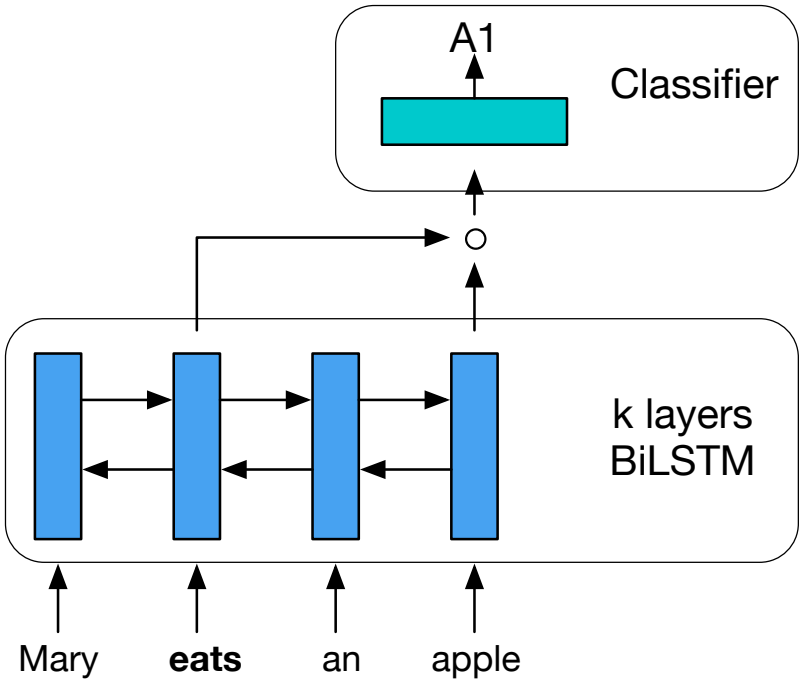


Figure 2: Predicting an argument and its label with an LSTM encoder.

[Marcheggiani et al., 2017]

Neural Semantic Parsing

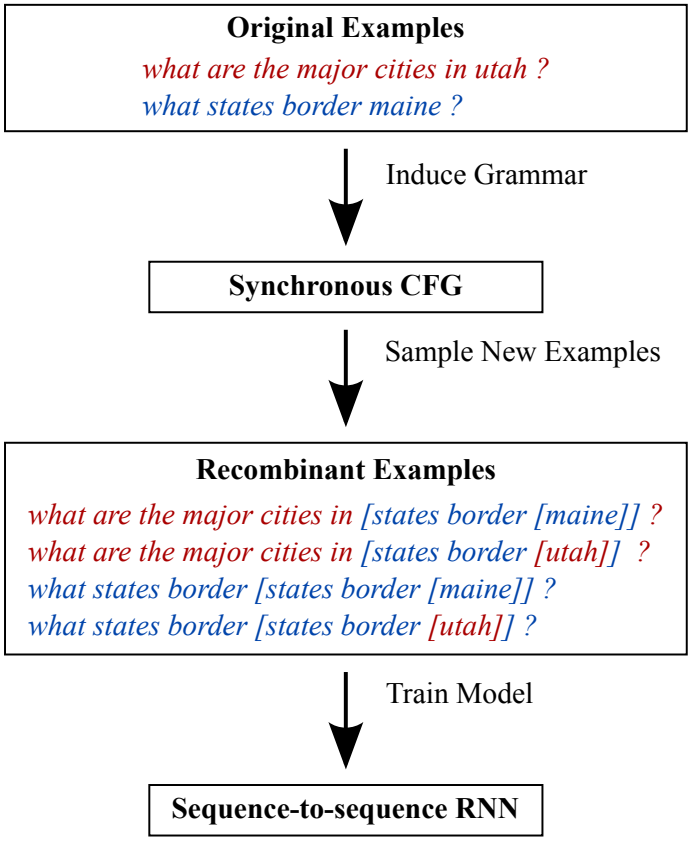


Figure 1: An overview of our system. Given a dataset, we induce a high-precision synchronous context-free grammar. We then sample from this grammar to generate new "recombinant" examples, which we use to train a sequence-to-sequence RNN.

GEO
x: "*what is the population of iowa ?*"
y:
```
_answer ( NV , (
  _population ( NV , V1 ) , _const (
    V0 , _stateid ( iowa ) ) ) )
```

ATIS
x: "*can you list all flights from chicago to milwaukee*"
y:
```
( _lambda $0 e ( _and
  ( _flight $0 )
  ( _from $0 chicago :  _ci )
  ( _to $0 milwaukee :  _ci ) ) )
```

Overnight
x: "*when is the weekly standup*"
y:
```
( call listValue ( call
    getProperty meeting.weekly_standup
    ( string start_time ) ) )
```

Figure 2: One example from each of our domains. We tokenize logical forms as shown, thereby casting semantic parsing as a sequence-to-sequence task.

[Jia and Liang, 2016]

Neural Semantic Parsing

Examples
("*what states border texas ?*",
`answer(NV, (state(V0), next_to(V0, NV), const(V0, stateid(texas))))`)
("*what is the highest mountain in ohio ?*",
`answer(NV, highest(V0, (mountain(V0), loc(V0, NV), const(V0, stateid(ohio)))))`)

Rules created by ABSENTITIES
ROOT → ⟨ "*what states border* STATEID *?*",
`answer(NV, (state(V0), next_to(V0, NV), const(V0, stateid(STATEID))))`⟩
STATEID → ⟨ "*texas*", `texas` ⟩
ROOT → ⟨ "*what is the highest mountain in* STATEID *?*",
`answer(NV, highest(V0, (mountain(V0), loc(V0, NV), const(V0, stateid(STATEID)))))`⟩
STATEID → ⟨"*ohio*", `ohio`⟩

Rules created by ABSWHOLEPHRASES
ROOT → ⟨ "*what states border* STATE *?*", `answer(NV, (state(V0), next_to(V0, NV), STATE))`⟩
STATE → ⟨ "*states border texas*", `state(V0), next_to(V0, NV), const(V0, stateid(texas))`⟩
ROOT → ⟨ "*what is the highest mountain in* STATE *?*",
`answer(NV, highest(V0, (mountain(V0), loc(V0, NV), STATE)))`⟩

Rules created by CONCAT-2
ROOT → ⟨SENT$_1$ `</s>` SENT$_2$, SENT$_1$ `</s>` SENT$_2$⟩
SENT → ⟨ "*what states border texas ?*",
`answer(NV, (state(V0), next_to(V0, NV), const(V0, stateid(texas))))` ⟩
SENT → ⟨ "*what is the highest mountain in ohio ?*",
`answer(NV, highest(V0, (mountain(V0), loc(V0, NV), const(V0, stateid(ohio)))))` ⟩

Figure 3: Various grammar induction strategies illustrated on GEO. Each strategy converts the rules of an input grammar into rules of an output grammar. This figure shows the base case where the input grammar has rules ROOT → $\langle x, y \rangle$ for each (x, y) pair in the training dataset.

[Jia and Liang, 2016]

Neural Semantic Parsing

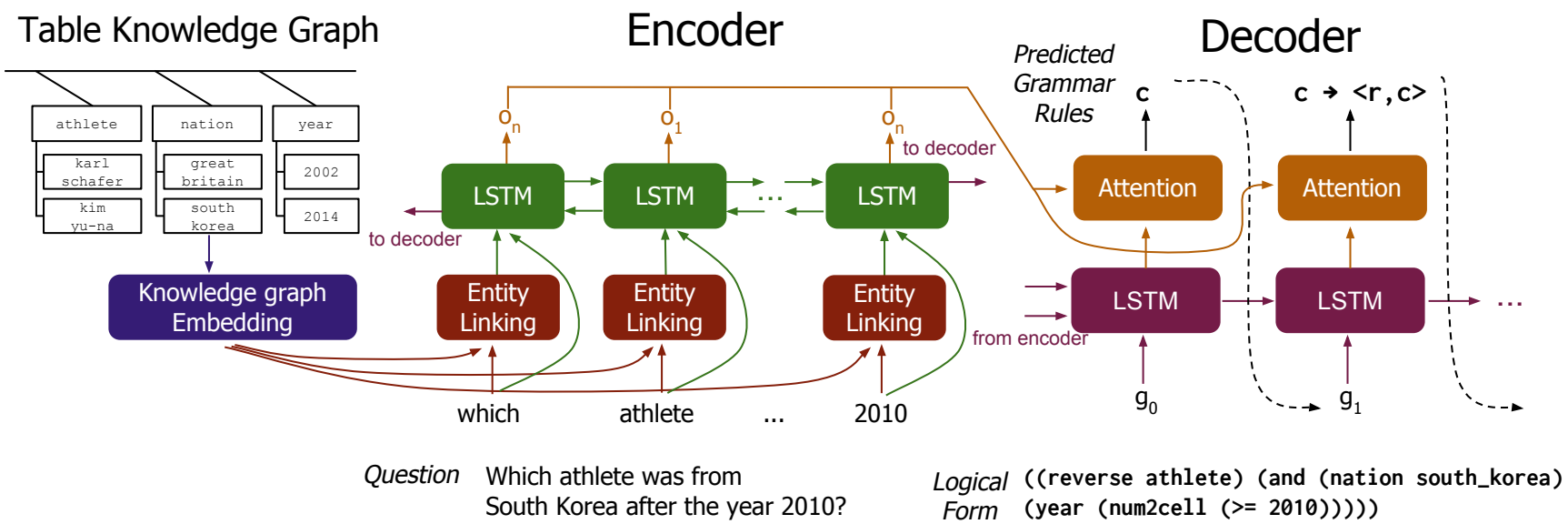
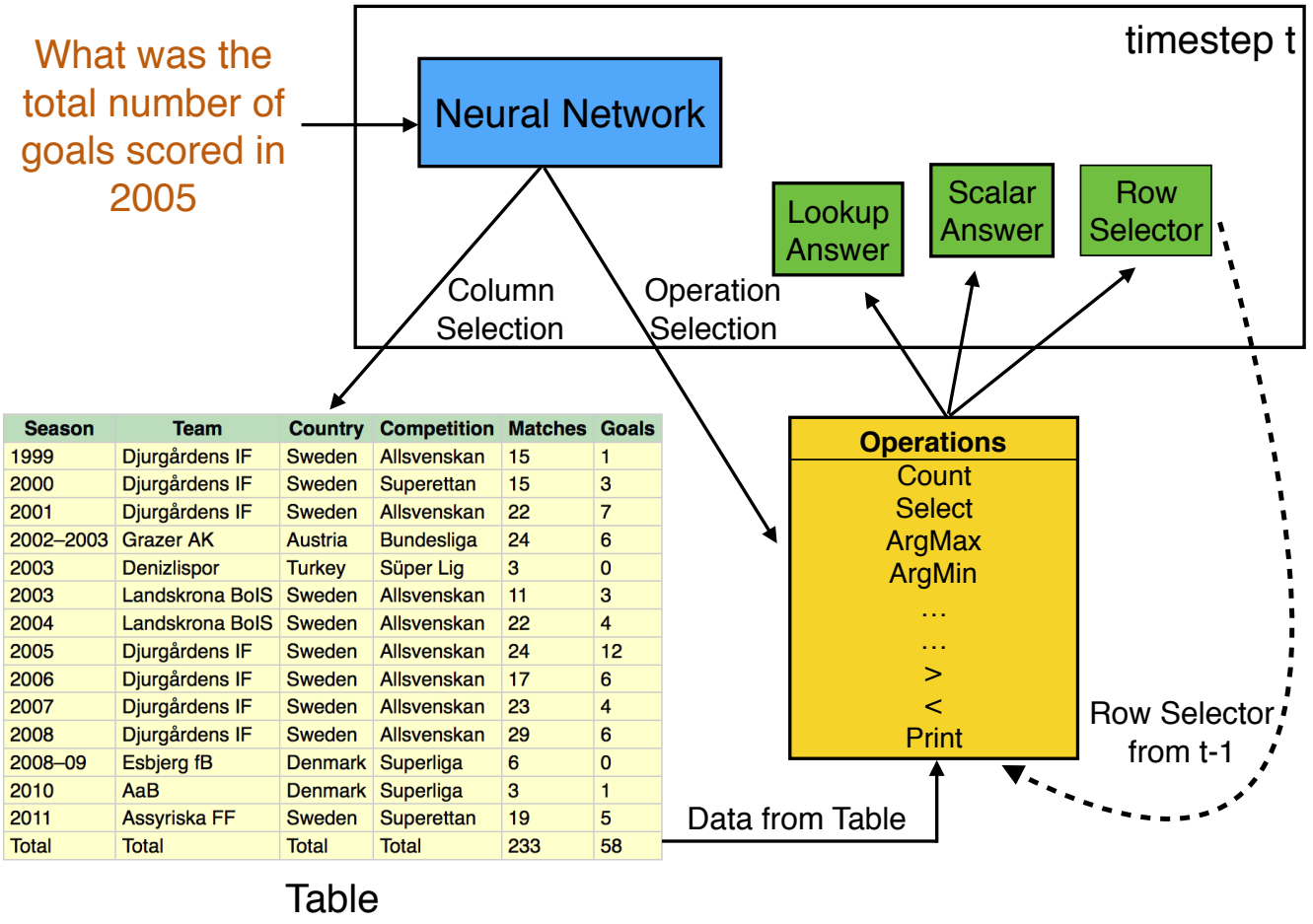


Figure 1: Overview of our semantic parsing model. The encoder performs entity embedding and linking before encoding the question with a bidirectional LSTM. The decoder predicts a sequence of grammar rules that generate a well-typed logical form.

[Krishnamurthy et al., 2017]

Neural Semantic Parsing



Season	Team	Country	Competition	Matches	Goals
1999	Djurgårdens IF	Sweden	Allsvenskan	15	1
2000	Djurgårdens IF	Sweden	Superettan	15	3
2001	Djurgårdens IF	Sweden	Allsvenskan	22	7
2002–2003	Grazer AK	Austria	Bundesliga	24	6
2003	Denizlispor	Turkey	Süper Lig	3	0
2003	Landskrona BoIS	Sweden	Allsvenskan	11	3
2004	Landskrona BoIS	Sweden	Allsvenskan	22	4
2005	Djurgårdens IF	Sweden	Allsvenskan	24	12
2006	Djurgårdens IF	Sweden	Allsvenskan	17	6
2007	Djurgårdens IF	Sweden	Allsvenskan	23	4
2008	Djurgårdens IF	Sweden	Allsvenskan	29	6
2008–09	Esbjerg fB	Denmark	Superliga	6	0
2010	AaB	Denmark	Superliga	3	1
2011	Assyriska FF	Sweden	Superettan	19	5
Total	Total	Total	Total	233	58

Figure 1: Neural Programmer is a neural network augmented with a set of discrete operations. The model runs for a fixed number of time steps, selecting an operation and a column from the table at every time step. The induced program transfers information across timesteps using the *row selector* variable while the output of the model is stored in the *scalar answer* and *lookup answer* variables.

[Neelakantan et al., 2017]

Neural AMR Parsing

Obama was elected and his voters celebrated

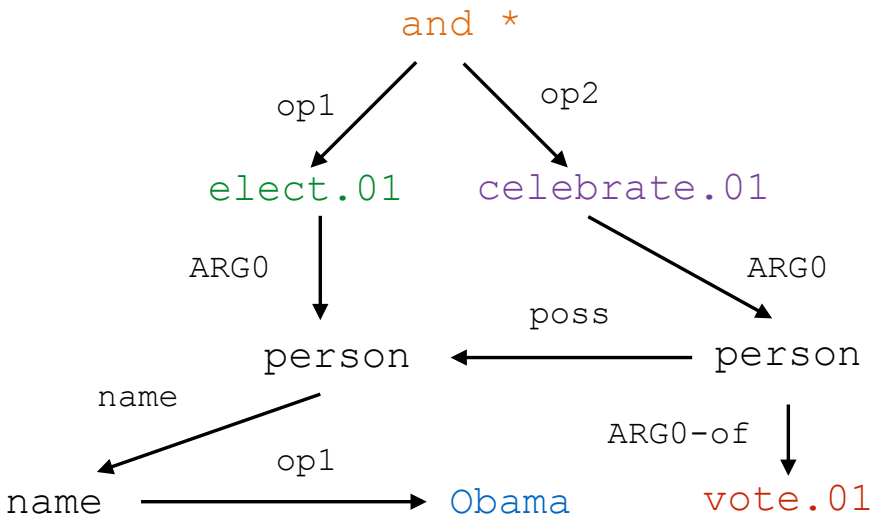
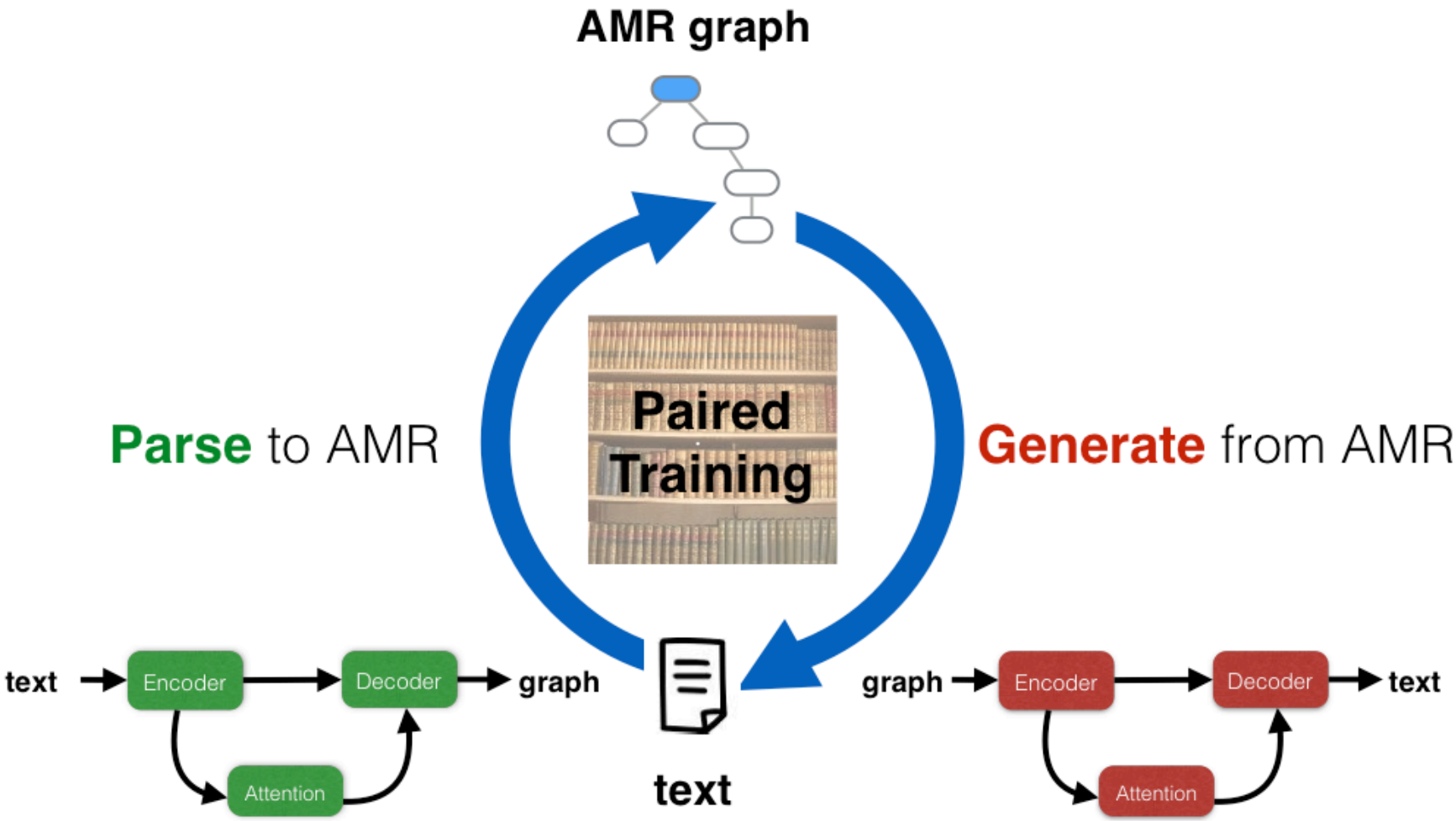


Figure 1: An example sentence and its corresponding Abstract Meaning Representation (AMR). AMR encodes semantic dependencies between entities mentioned in the sentence, such as "Obama" being the "arg0" of the verb "elected".

[Konstas et al., 2017]

Neural AMR Parsing



[Konstas et al., 2017]

Question Answering

IR-based Question Answering

- Initial approaches to Q&A: pattern matching, pattern learning, query rewriting, information extraction

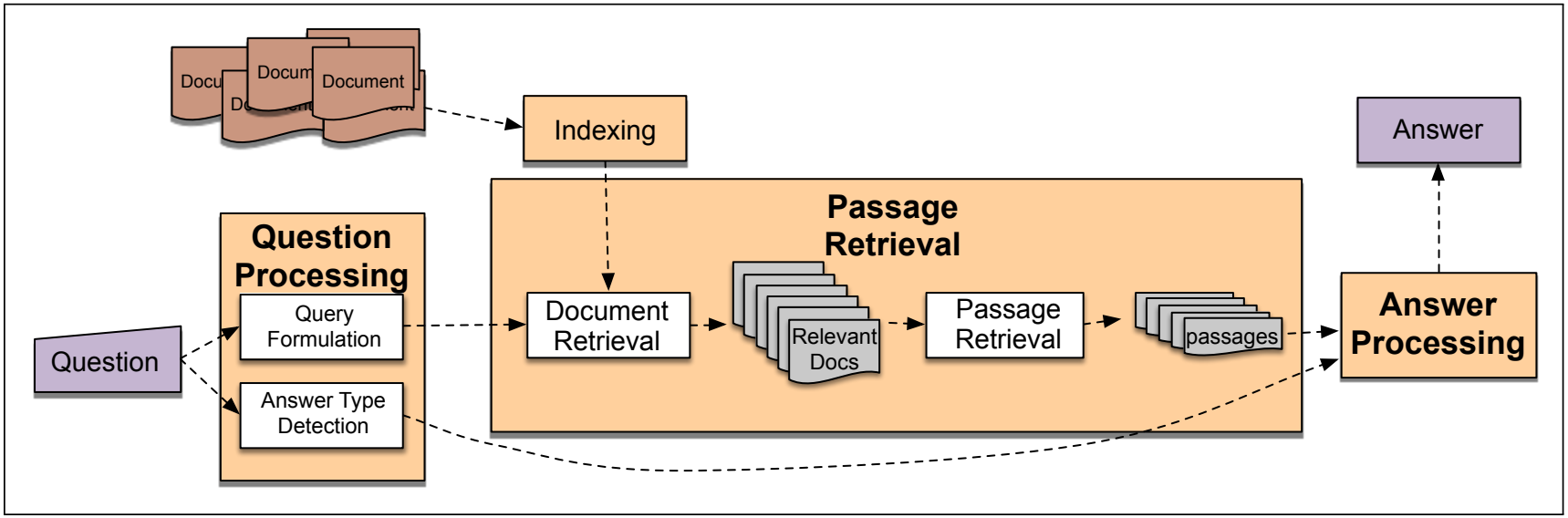


Figure 28.2 IR-based factoid question answering has three stages: question processing, passage retrieval, and answer processing.

[Jurafsky-SLP3]

IR-based Question Answering

- Large-scale, open-domain IE system like IBM Watson

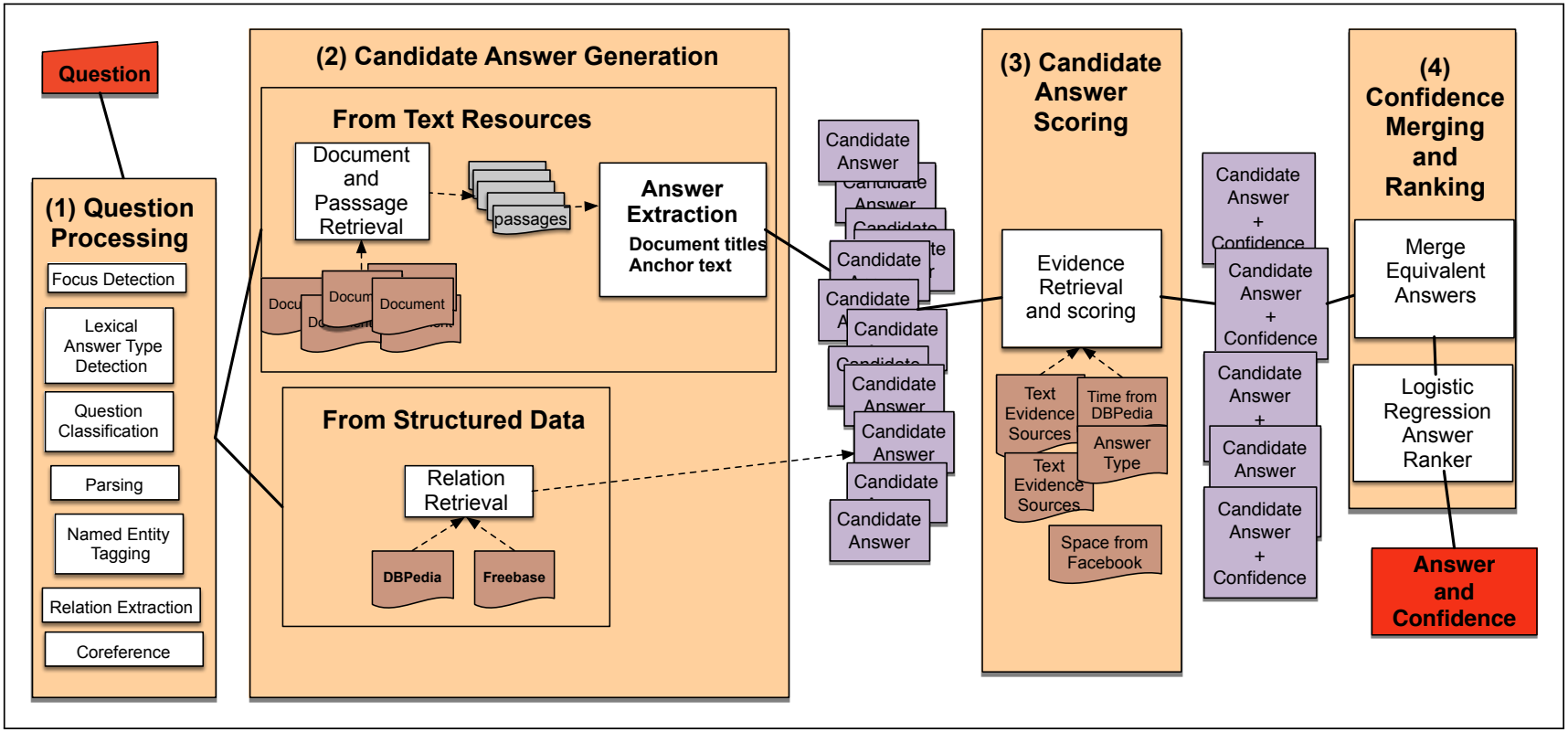
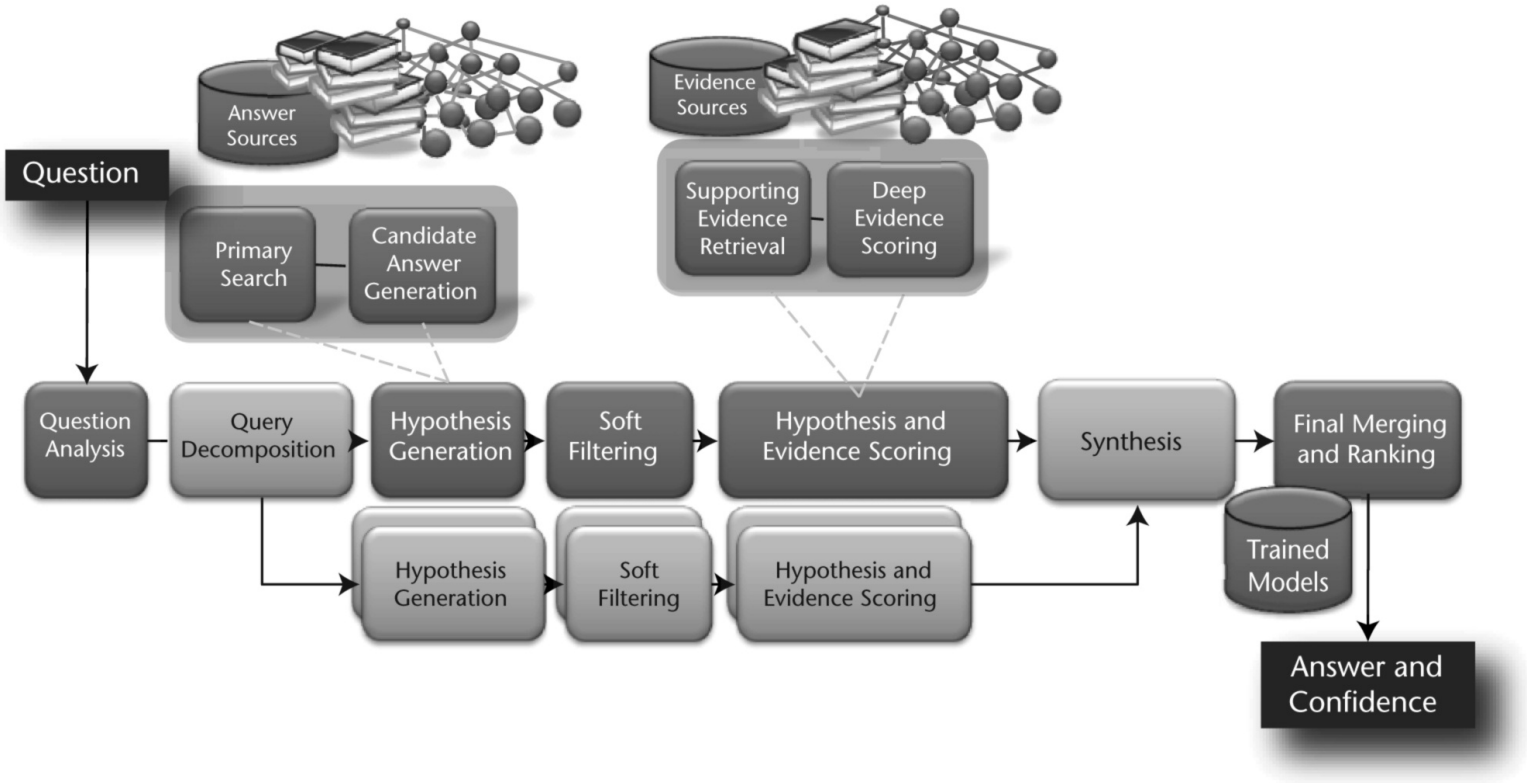


Figure 28.9 The 4 broad stages of Watson QA: (1) Question Processing, (2) Candidate Answer Generation, (3) Candidate Answer Scoring, and (4) Answer Merging and Confidence Scoring.

[Jurafsky-SLP3]

IR-based Question Answering

- Large-scale, open-domain IE system like IBM Watson



[Ferrucci et al., 2010]

IR-based Question Answering

- Answer types:

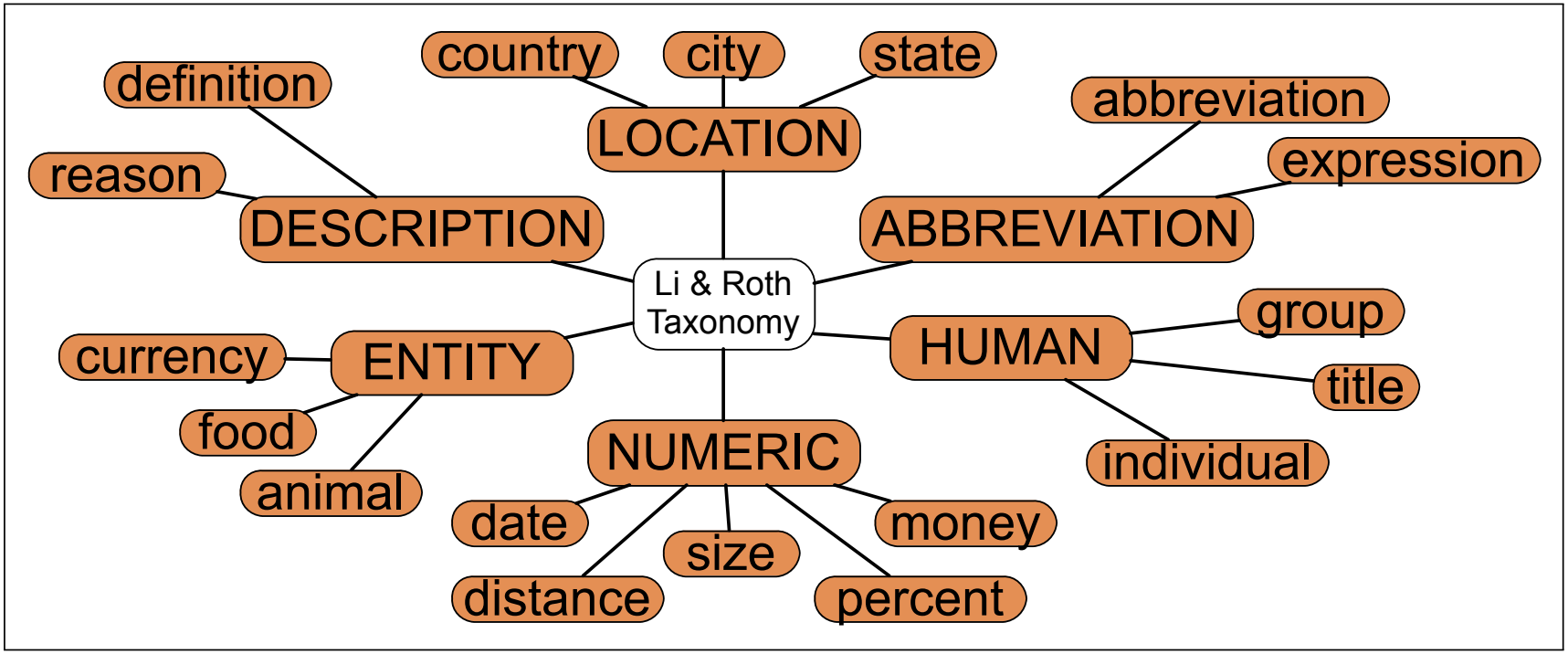


Figure 28.3 A subset of the Li and Roth (2005) answer types.

[Jurafsky-SLP3]

IR-based Question Answering

- Query types:

HUMAN	
description	Who was Confucius?
group	What are the major companies that are part of Dow Jones?
ind	Who was the first Russian astronaut to do a spacewalk?
title	What was Queen Victoria's title regarding India?
LOCATION	
city	What's the oldest capital city in the Americas?
country	What country borders the most others?
mountain	What is the highest peak in Africa?
other	What river runs through Liverpool?
state	What states do not have state income tax?
NUMERIC	
code	What is the telephone number for the University of Colorado?
count	About how many soldiers died in World War II?
date	What is the date of Boxing Day?
distance	How long was Mao's 1930s Long March?
money	How much did a McDonald's hamburger cost in 1963?
order	Where does Shanghai rank among world cities in population?
other	What is the population of Mexico?
period	What was the average life expectancy during the Stone Age?
percent	What fraction of a beaver's life is spent swimming?
temp	How hot should the oven be when making Peachy Oat Muffins?
speed	How fast must a spacecraft travel to escape Earth's gravity?
size	What is the size of Argentina?
weight	How many pounds are there in a stone?

Figure 25.4 Question typology from Li and Roth (2002), (2005). Example sentences are from their corpus of 5500 labeled questions. A question can be labeled either with a coarse-grained tag like HUMAN or NUMERIC or with a fine-grained tag like HUMAN:DESCRIPTION, HUMAN:GROUP, HUMAN:IND, and so on.

[Jurafsky-SLP3]

IR-based Question Answering

- Query types:

Tag	Example
ABBREVIATION	
abb	What's the abbreviation for limited partnership?
exp	What does the "c" stand for in the equation E=mc2?
DESCRIPTION	
definition	What are tannins?
description	What are the words to the Canadian National anthem?
manner	How can you get rust stains out of clothing?
reason	What caused the Titanic to sink?
ENTITY	
animal	What are the names of Odin's ravens?
body	What part of your body contains the corpus callosum?
color	What colors make up a rainbow?
creative	In what book can I find the story of Aladdin?
currency	What currency is used in China?
disease/medicine	What does Salk vaccine prevent?
event	What war involved the battle of Chapultepec?
food	What kind of nuts are used in marzipan?
instrument	What instrument does Max Roach play?
lang	What's the official language of Algeria?
letter	What letter appears on the cold-water tap in Spain?
other	What is the name of King Arthur's sword?
plant	What are some fragrant white climbing roses?
product	What is the fastest computer?
religion	What religion has the most members?
sport	What was the name of the ball game played by the Mayans?
substance	What fuel do airplanes use?
symbol	What is the chemical symbol for nitrogen?
technique	What is the best way to remove wallpaper?
term	How do you say " Grandma" in Irish?
vehicle	What was the name of Captain Bligh's ship?
word	What's the singular of dice?

[Jurafsky-SLP3]

Answer Extraction

25.1.6 Feature-based Answer Extraction

Supervised learning approaches to answer extraction train classifiers to decide if a span or a sentence contains an answer. One obviously useful feature is the answer type feature of the above baseline algorithm. Hand-written regular expression patterns also play a role, such as the sample patterns for definition questions in Fig. 25.5.

Pattern	Question	Answer
<AP> such as <QP>	What is autism?	", developmental disorders such as autism"
<QP>, a <AP>	What is a caldera?	"the Long Valley caldera, a volcanic crater 19 miles long"

Figure 25.5 Some answer-extraction patterns using the answer phrase (AP) and question phrase (QP) for definition questions (Pasca, 2003).

Other features in such classifiers include:

Answer type match: True if the candidate answer contains a phrase with the correct answer type.

Pattern match: The identity of a pattern that matches the candidate answer.

Number of matched question keywords: How many question keywords are contained in the candidate answer.

Keyword distance: The distance between the candidate answer and query keywords.

Novelty factor: True if at least one word in the candidate answer is novel, that is, not in the query.

Apposition features: True if the candidate answer is an appositive to a phrase containing many question terms. Can be approximated by the number of question terms separated from the candidate answer through at most three words and one comma (Pasca, 2003).

[Jurafsky-SLP3]

Neural Answer Extraction

Beyoncé Giselle Knowles-Carter (born September 4, 1981) is an American singer, songwriter, record producer and actress. Born and raised in Houston, Texas, she performed in various singing and dancing competitions as a child, and rose to fame in the late 1990s as lead singer of R&B girl-group Destiny's Child. Managed by her father, Mathew Knowles, the group became one of the world's best-selling girl groups of all time. Their hiatus saw the release of Beyoncé's debut album, Dangerously in Love (2003), which established her as a solo artist worldwide, earned five Grammy Awards and featured the Billboard Hot 100 number-one singles "Crazy in Love" and "Baby Boy".
Q: "In what city and state did Beyoncé grow up?" A: "Houston, Texas"
Q: "What areas did Beyoncé compete in when she was growing up?" A: "singing and dancing"
Q: "When did Beyoncé release Dangerously in Love?" A: "2003"

Figure 25.6 A (Wikipedia) passage from the SQuAD 2.0 dataset (Rajpurkar et al., 2018) with 3 sample questions and the labeled answer spans.

[Jurafsky-SLP3]

Passage-based Q&A

James the Turtle was always getting in trouble. Sometimes he'd reach into the freezer and empty out all the food. Other times he'd sled on the deck and get a splinter. … He went to the grocery store and pulled all the pudding off the shelves and ate two jars. Then he walked to the fast food restaurant and ordered 15 bags of fries.

Q. What did James pull off of the shelves in the grocery store?
(A) pudding, (B) fries, (C) food, (D) splinters

Q. Where did James go after eating two jars of pudding?
(A) grocery, (B) restaurant, (C) freezer, (D) home

CNN/DailyMail RC Datasets

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]

CNN/DailyMail RC Datasets

	CNN			Daily Mail		
	train	valid	test	train	valid	test
# months	95	1	1	56	1	1
# documents	90,266	1,220	1,093	196,961	12,148	10,397
# queries	380,298	3,924	3,198	879,450	64,835	53,182
Max # entities	527	187	396	371	232	245
Avg # entities	26.4	26.5	24.5	26.5	25.5	26.0
Avg # tokens	762	763	716	813	774	780
Vocab size	118,497			208,045		

[Hermann et al. 2015]

Who-did-What (WDW) Dataset

- Solves several issues with CNN/DM dataset:
  - Starts with the selection of a question article from Gigaword corpus
  - Question is formed by deleting a person named entity from the first sentence of the question article
  - An information retrieval system is then used to select a passage with high overlap with the first sentence of the question article, and an answer choice list is generated from the person named entities in the passage

  - Forms questions from two distinct articles rather than summary points
  - Allows using documents that don't contain manually-written summaries
  - Reduces syntactic similarity between question & relevant passage sentences
  - Selectively remove problems so as to suppress four simple baselines — selecting the most mentioned person, the first mentioned person, and two language model baselines

  - The resulting dataset yields a larger gap between human and machine performance than existing ones, i.e., humans can answer more questions, while existing state-of-the-art models perform worse!

[Onishi et al. 2016]

Who-did-What (WDW) Dataset

Passage: Britain's decision on Thursday to drop extradition proceedings against Gen. Augusto Pinochet and allow him to return to Chile is understandably frustrating ... Jack Straw, the home secretary, said the 84-year-old former dictator's ability to understand the charges against him and to direct his defense had been seriously impaired by a series of strokes. ... Chile's president-elect, Ricardo Lagos, has wisely pledged to let justice run its course. But the outgoing government of President Eduardo Frei is pushing a constitutional reform that would allow Pinochet to step down from the Senate and retain parliamentary immunity from prosecution. ...

Question: Sources close to the presidential palace said that Fujimori declined at the last moment to leave the country and instead he will send a high level delegation to the ceremony, at which Chilean President Eduardo Frei will pass the mandate to XXX.

Choices: (1) Augusto Pinochet (2) Jack Straw (3) Ricardo Lagos

Passage: Tottenham won 2-0 at Hapoel Tel Aviv in UEFA Cup action on Thursday night in a defensive display which impressed Spurs skipper Robbie Keane. ... Keane scored the first goal at the Bloomfield Stadium with Dimitar Berbatov, who insisted earlier on Thursday he was happy at the London club, heading a second. The 26-year-old Berbatov admitted the reports linking him with a move had affected his performances ... Spurs manager Juande Ramos has won the UEFA Cup in the last two seasons ...

Question: Tottenham manager Juande Ramos has hinted he will allow XXX to leave if the Bulgaria striker makes it clear he is unhappy.

Choices: (1) Robbie Keane (2) Dimitar Berbatov

[Onishi et al. 2016]

Who-did-What (WDW) Dataset

Baseline	Accuracy	
	Before	After
First person in passage	0.60	0.32
Most frequent person	0.61	0.33
n-gram	0.53	0.33
Unigram	0.43	0.32
Random*	0.32	0.32

Table 2: Performance of suppressed baselines. *Random performance is computed as a deterministic function of the number of times each choice set size appears. Many questions have only two choices and there are about three choices on average.

	relaxed train	train	valid	test
# queries	185,978	127,786	10,000	10,000
Avg # choices	3.5	3.5	3.4	3.4
Avg # tokens	378	365	325	326
Vocab size	347,406		308,602	

Table 3: Dataset statistics.

[Onishi et al. 2016]

Who-did-What (WDW) Dataset

System	WDW	CNN
Word overlap	0.47	–
Sliding window	0.48	–
Distance	0.46	–
Sliding window + Distance	0.51	–
Semantic features	0.52	–
Attentive Reader	0.53	0.63^{I}
Attentive Reader (relaxed train)	0.55	
Stanford Reader	0.64	0.73^{II}
Stanford Reader (relaxed train)	0.65	
AS Reader	0.57	0.70^{III}
AS Reader (relaxed train)	0.59	
GA Reader	0.57	0.74^{IV}
GA Reader (relaxed train)	0.60	
Human Performance	$84/100$	$0.75+^{II}$

[Onishi et al. 2016]

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]

SQuAd Dataset (100K Manually-Labeled)

Paragraph 1 of 43

Spend around 4 minutes on the following paragraph to ask 5 questions! If you can't ask 5 questions, ask 4 or 3 (worse), but do your best to ask 5. Select the answer from the paragraph by clicking on 'Select Answer', and then highlight the smallest segment of the paragraph that answers the question.

Oxygen is a chemical element with symbol O and atomic number 8. It is a member of the chalcogen group on the periodic table and is a highly reactive nonmetal and oxidizing agent that readily forms compounds (notably oxides) with most elements. By mass, oxygen is the third-most abundant element in the universe, after hydrogen and helium. At standard temperature and pressure, two atoms of the element bind to form dioxygen, a colorless and odorless diatomic gas with the formula O
2. Diatomic oxygen gas constitutes 20.8% of the Earth's atmosphere. However, monitoring of atmospheric oxygen levels show a global downward trend, because of fossil-fuel burning. Oxygen is the most abundant element by mass in the Earth's crust as part of oxide compounds such as silicon dioxide, making up almost half of the crust's mass.

When asking questions, **avoid using** the same words/phrases as in the paragraph. Also, you are encouraged to pose **hard questions**.

Ask a question here. Try using your own words

Select Answer

Ask a question here. Try using your own words

Select Answer

[Rajpurkar et al. 2016]

SQuAd Dataset (100K Manually-Labeled)

Reasoning	Description	Example	Percentage
Lexical variation (synonymy)	Major correspondences between the question and the answer sentence are synonyms.	Q: What is the Rankine cycle sometimes **called**? Sentence: The Rankine cycle is sometimes **referred** to as a practical Carnot cycle.	33.3%
Lexical variation (world knowledge)	Major correspondences between the question and the answer sentence require world knowledge to resolve.	Q: Which **governing bodies** have veto power? Sen.: **The European Parliament and the Council of the European Union** have powers of amendment and veto during the legislative process.	9.1%
Syntactic variation	After the question is paraphrased into declarative form, its syntactic dependency structure does not match that of the answer sentence even after local modifications.	Q: What Shakespeare scholar **is currently on the faculty**? Sen.: **Current faculty include** the anthropologist Marshall Sahlins, ..., Shakespeare scholar David Bevington.	64.1%
Multiple sentence reasoning	There is anaphora, or higher-level fusion of multiple sentences is required.	Q: What collection does **the V&A Theatre & Performance galleries** hold? Sen.: **The V&A Theatre & Performance galleries** opened in March 2009. ... **They** hold the UK's biggest national collection of material about live performance.	13.6%
Ambiguous	We don't agree with the crowdworkers' answer, or the question does not have a unique answer.	Q: What is the main goal of criminal punishment? Sen.: **Achieving crime control via incapacitation and deterrence** is a major goal of criminal punishment.	6.1%

Table 3: We manually labeled 192 examples into one or more of the above categories. Words relevant to the corresponding reasoning type are bolded, and the crowdsourced answer is underlined.

[Rajpurkar et al. 2016]

Facebook bAbI Tasks (Synthetic)

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 (Synthetic)

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

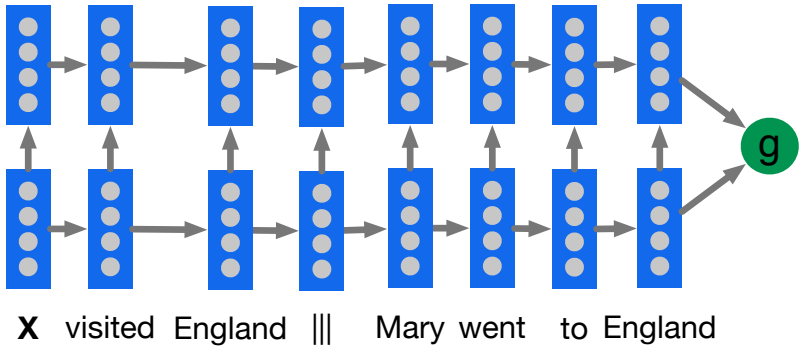
[Weston et al. 2016]

Facebook bAbI Tasks (Synthetic)

	Weakly Supervised		Uses External Resources	Strong Supervision (using supporting facts)						
TASK	N-gram Classifier	LSTM	Structured SVM *COREF+SRL features*	MemNN Weston et al. (2014)	MemNN *ADAPTIVE MEMORY*	MemNN *AM + N-GRAMS*	MemNN *AM + NONLINEAR*	MemNN *AM + NG + NL*	No. of ex. req. ≥ 95	MultiTask Training
1 - Single Supporting Fact	36	50	99	100	100	100	100	100	250 ex.	100
2 - Two Supporting Facts	2	20	74	100	100	100	100	100	500 ex.	100
3 - Three Supporting Facts	7	20	17	20	**100**	**99**	**100**	**100**	500 ex.	**98**
4 - Two Arg. Relations	50	61	98	71	69	**100**	73	**100**	500 ex.	80
5 - Three Arg. Relations	20	70	83	83	83	86	86	**98**	1000 ex.	**99**
6 - Yes/No Questions	49	48	99	47	52	53	**100**	**100**	500 ex.	**100**
7 - Counting	52	49	69	68	78	86	83	85	FAIL	86
8 - Lists/Sets	40	45	70	77	90	88	94	91	FAIL	93
9 - Simple Negation	62	64	100	65	71	63	**100**	**100**	500 ex.	**100**
10 - Indefinite Knowledge	45	44	99	59	57	54	**97**	**98**	1000 ex.	**98**
11 - Basic Coreference	29	72	100	100	100	100	100	100	250 ex.	100
12 - Conjunction	9	74	96	100	100	100	100	100	250 ex.	100
13 - Compound Coref.	26	94	99	100	100	100	100	100	250 ex.	100
14 - Time Reasoning	19	27	99	99	100	99	100	99	500 ex.	99
15 - Basic Deduction	20	21	96	74	73	**100**	77	**100**	100 ex.	**100**
16 - Basic Induction	43	23	24	27	**100**	**100**	**100**	**100**	100 ex.	94
17 - Positional Reasoning	46	51	61	54	46	49	57	65	FAIL	72
18 - Size Reasoning	52	52	62	57	50	74	54	**95**	1000 ex.	93
19 - Path Finding	0	8	49	0	9	3	15	36	FAIL	19
20 - Agent's Motivations	76	91	95	100	100	100	100	100	250 ex.	100
Mean Performance	34	49	79	75	79	83	87	93		92

[Weston et al. 2016]

Attentive/Impatient Readers

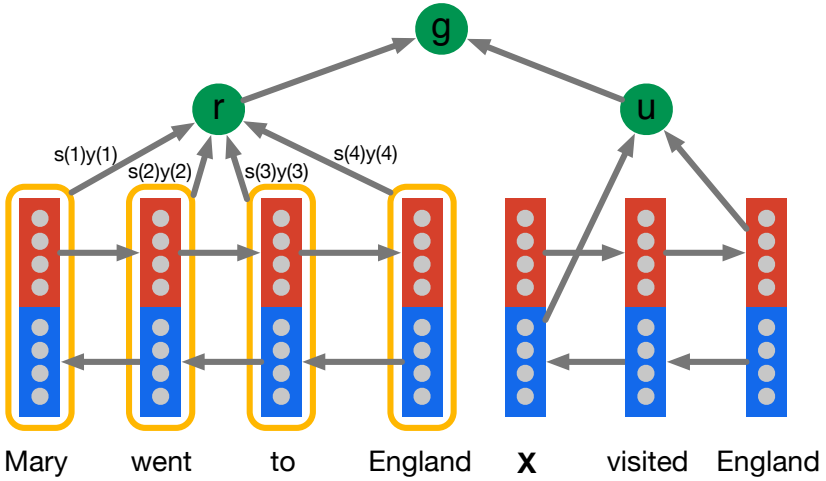


(c) A two layer Deep LSTM Reader with the question encoded before the document.

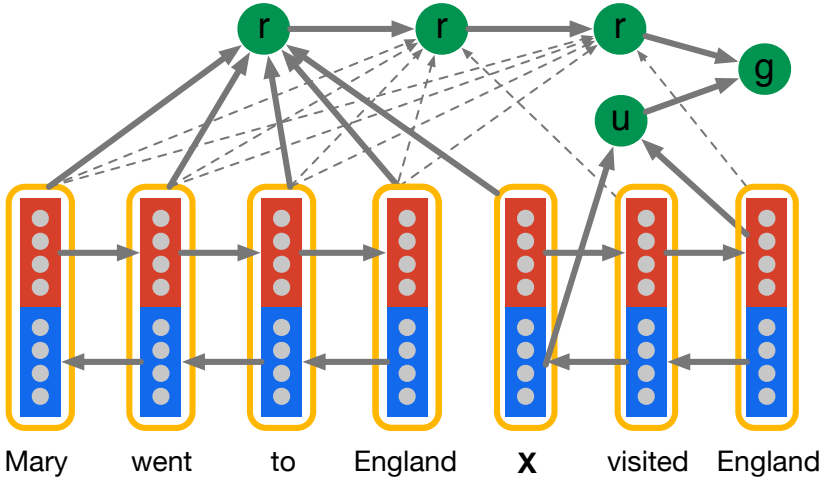
We feed our documents one word at a time into a Deep LSTM encoder, after a delimiter we then also feed the query into the encoder. Alternatively we also experiment with processing the query then the document. The result is that this model processes each document query pair as a single long sequence. Given the embedded document and query the network predicts which token in the document answers the query.

[Hermann et al. 2015]

Attentive/Impatient Readers



(a) Attentive Reader.



(b) Impatient Reader.

$$g^{\text{AR}}(d, q) = \tanh\left(W_{rg} r + W_{ug} u\right)$$

Ability to reread from the document as each query token is read. The result is an attention mechanism that allows the model to recurrently accumulate information from the document as it sees each query token, ultimately outputting a final joint document query representation for the answer prediction.

[Hermann et al. 2015]

Attentive/Impatient Readers

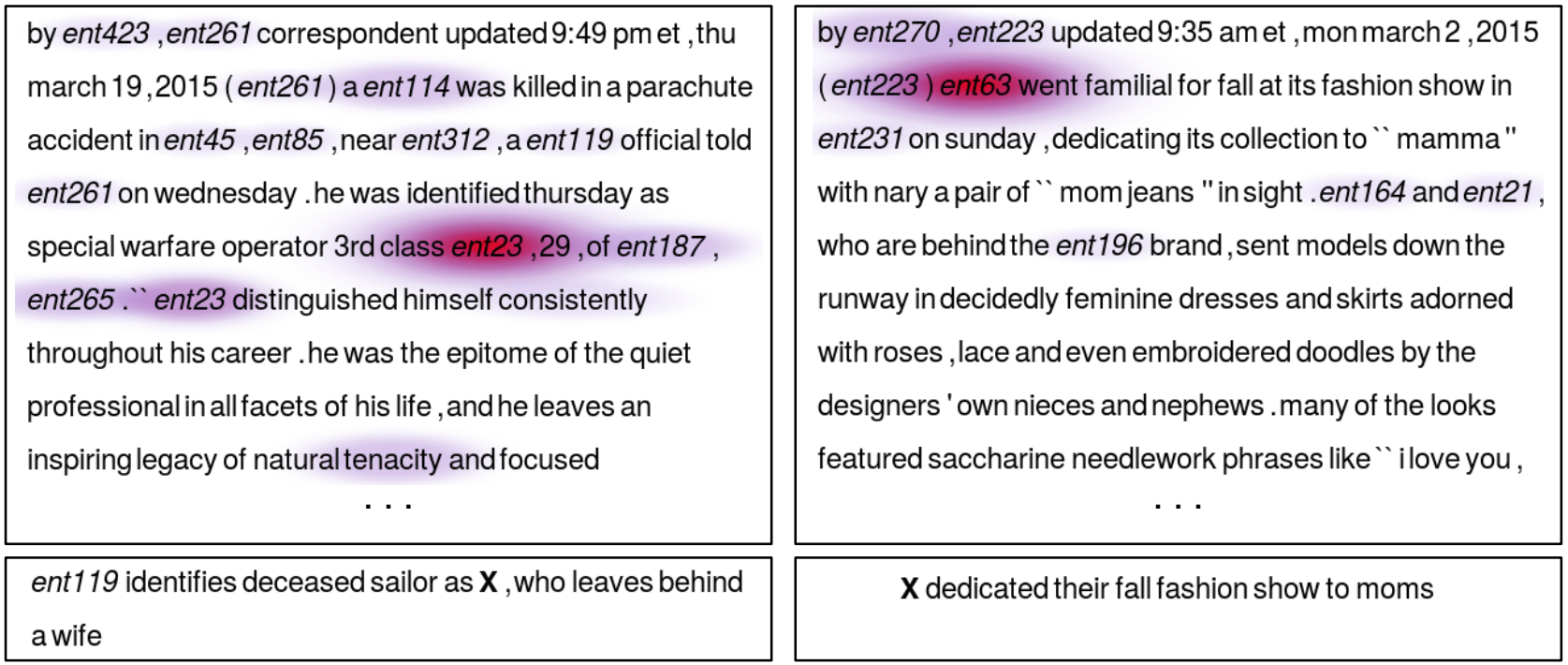


Figure 3: Attention heat maps from the Attentive Reader for two correctly answered validation set queries (the correct answers are *ent23* and *ent63*, respectively). Both examples require significant lexical generalisation and co-reference resolution in order to be answered correctly by a given model.

[Hermann et al. 2015]

Attentive/Impatient Readers

	CNN		Daily Mail	
	valid	test	valid	test
Maximum frequency	30.5	33.2	25.6	25.5
Exclusive frequency	36.6	39.3	32.7	32.8
Frame-semantic model	36.3	40.2	35.5	35.5
Word distance model	50.5	50.9	56.4	55.5
Deep LSTM Reader	55.0	57.0	63.3	62.2
Uniform Reader	39.0	39.4	34.6	34.4
Attentive Reader	61.6	63.0	**70.5**	**69.0**
Impatient Reader	**61.8**	**63.8**	69.0	68.0

[Hermann et al. 2015]

bi-LSTM-based Reading Comprehension Algo

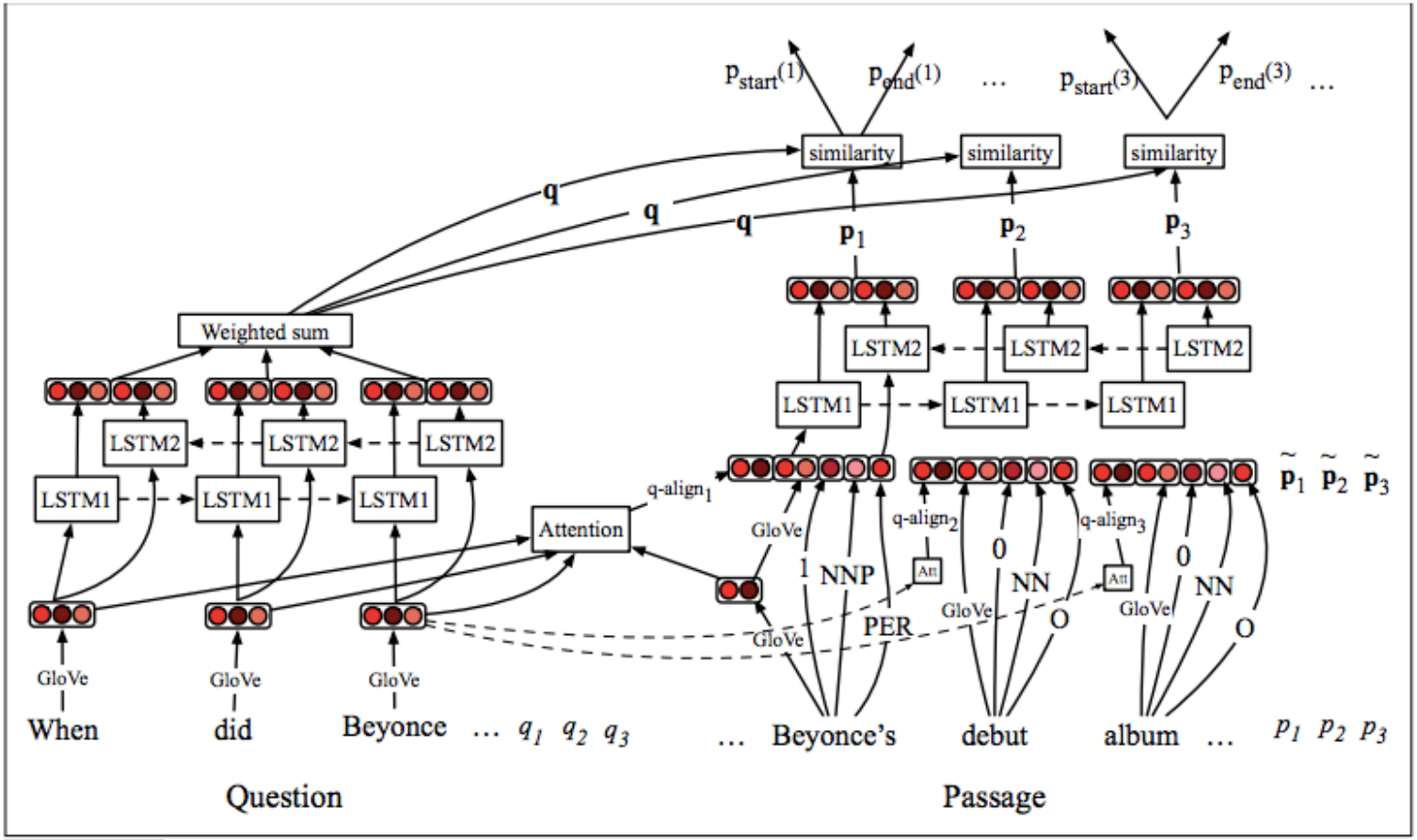


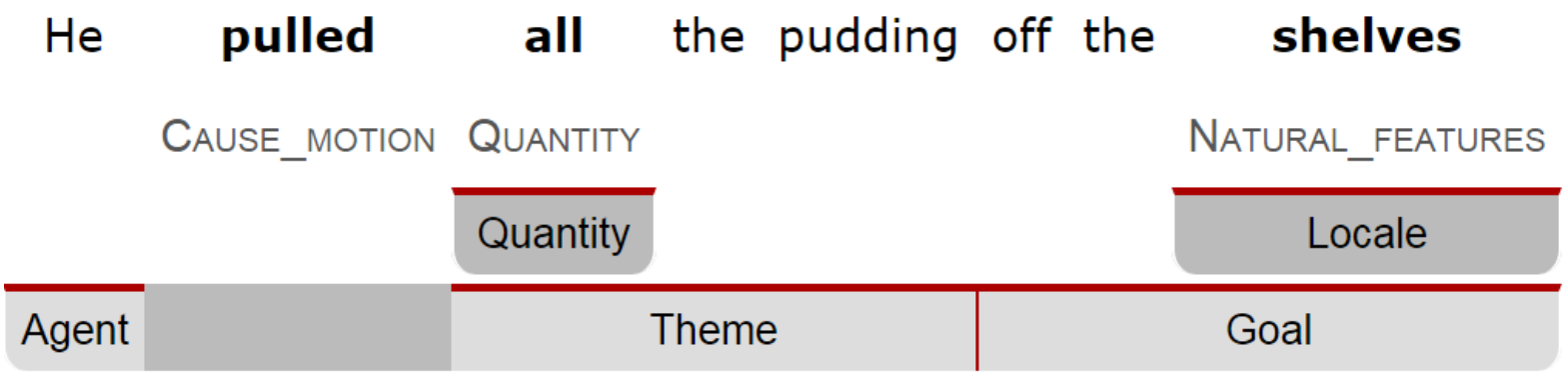
Figure 25.7 The question answering system of Chen et al. (2017), considering part of the question *When did Beyoncé release Dangerously in Love?* and the passage starting *Beyoncé's debut album, Dangerously in Love (2003).*

Like most such systems, DrQA builds an embedding for the question, builds an embedding for each token in the passage, computes a similarity function between the question and each passage word in context, and then uses the question-passage similarity scores to decide where the answer span starts and ends.

[Chen et al., 2017]

Feature-based Model

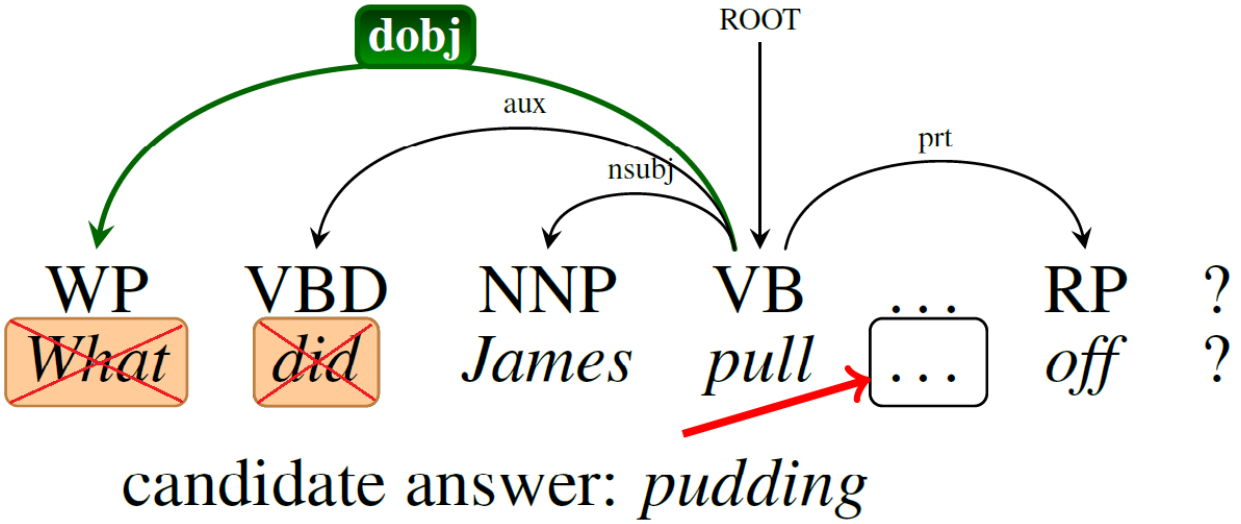
- Weighted word overlap between the bag of words constructed from the question/answer and in the window (and their word embedding versions)
- Minimal distance between two word occurrences in the passage that are also contained in the question/answer pair
- Frame semantics (predicates, frames evoked,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



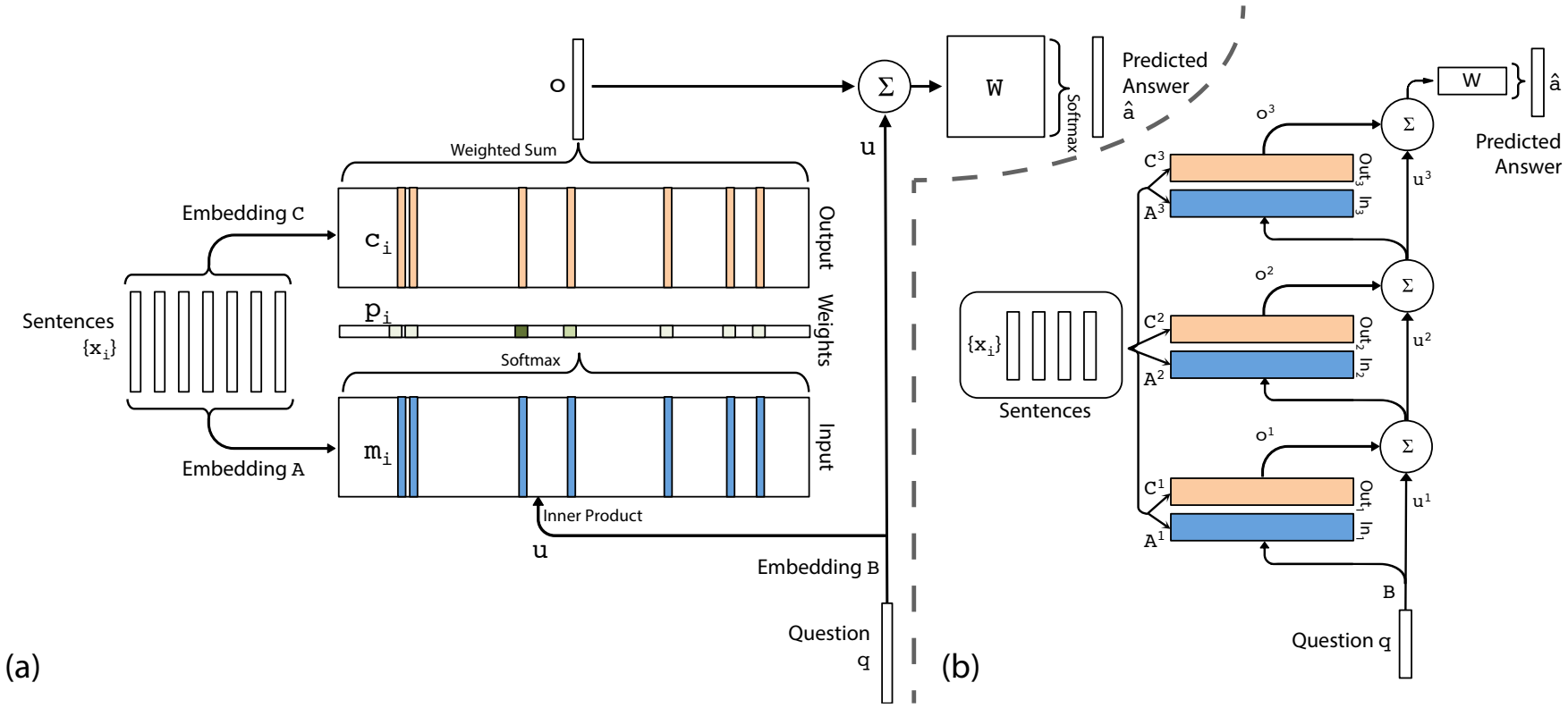
[Wang et al. 2015]

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

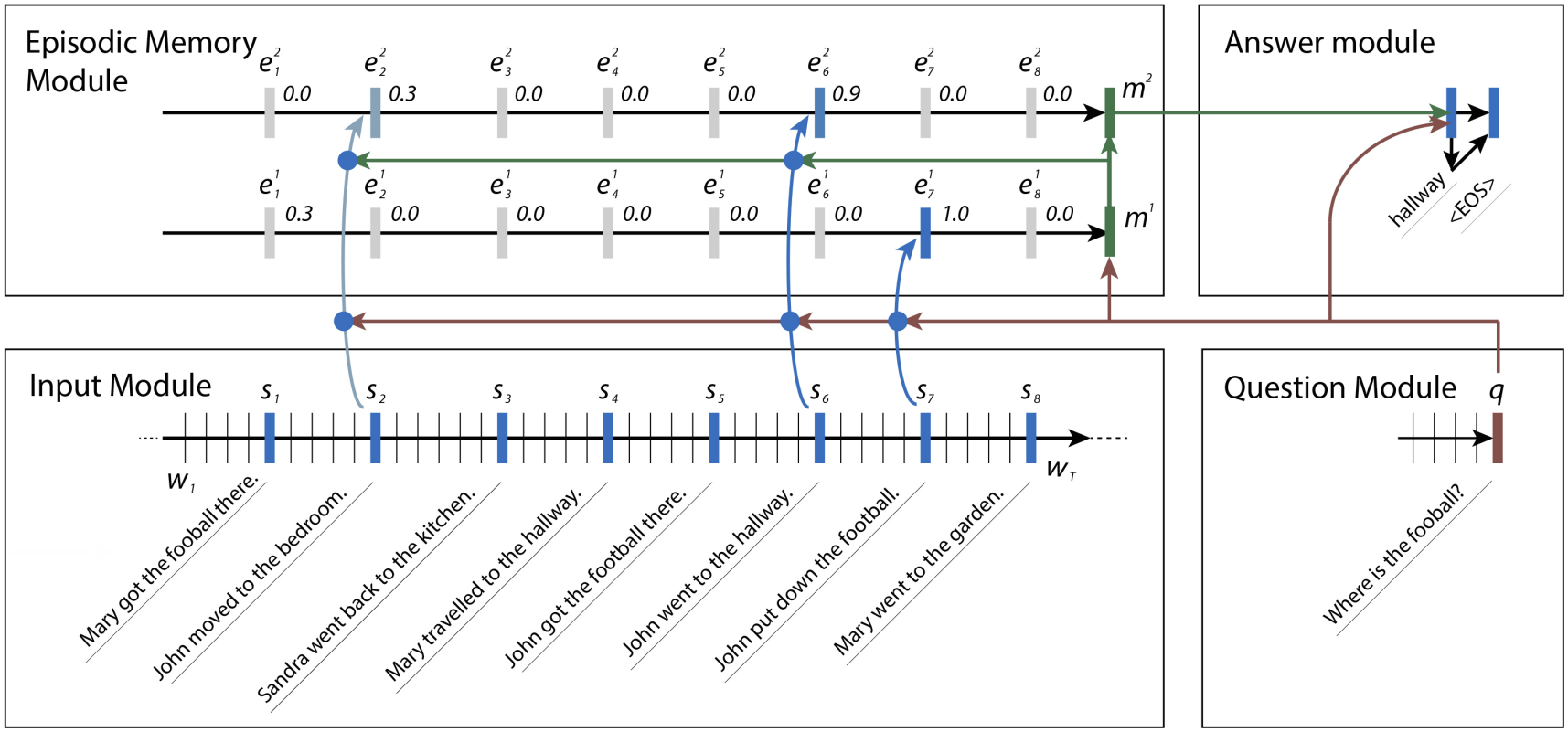
[Weston et al., 2014; Sukhbaatar et al., 2015; Kumar et al., 2016; Dhingra et al. 2017]

End-to-End Memory Networks



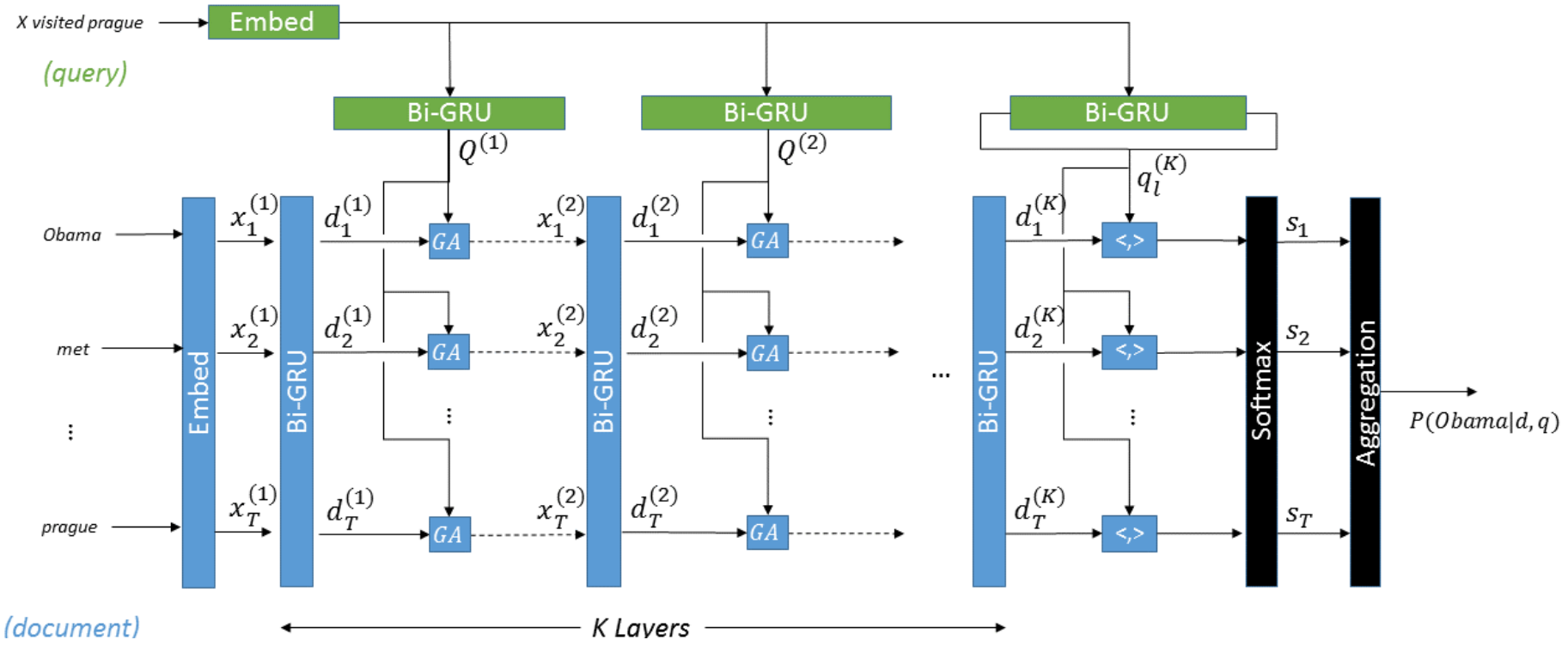
[Sukhbaatar et al., 2015]

Dynamic Memory Networks



[Kumar et al., 2016]

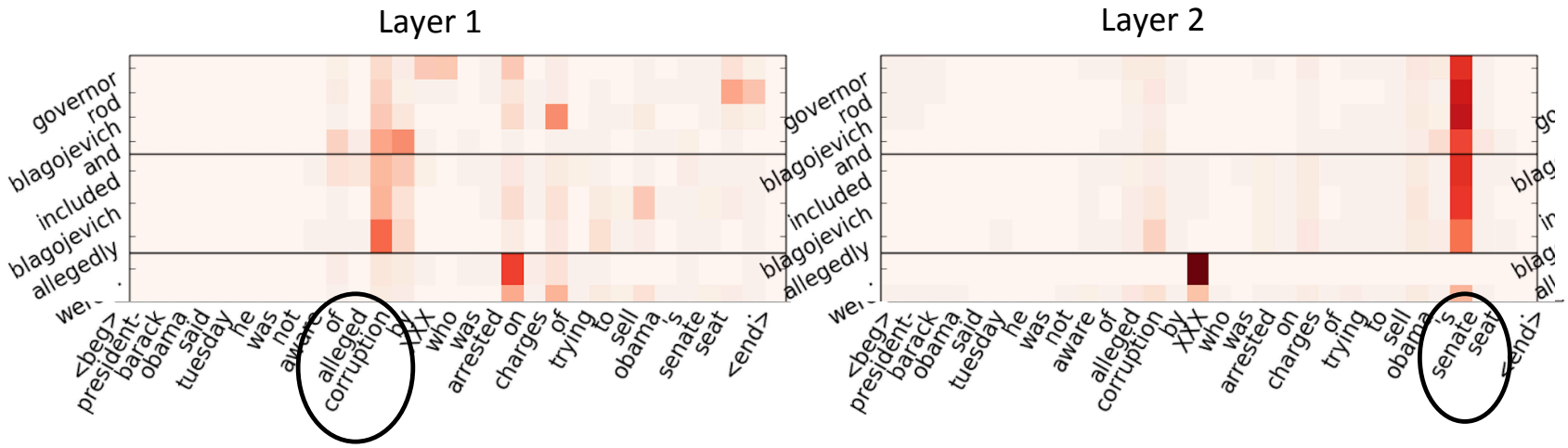
Gated Attention Readers



[Dhingra et al. 2017]

Gated Attention Readers

- **Context**: "…arrested Illinois governor **Rod Blagojevich** and his chief of staff John Harris on corruption charges … included **Blogojevich** allegedly conspiring to sell or trade the senate seat left vacant by President-elect Barack Obama…"
- **Query**: "President-elect Barack Obama said Tuesday he was not aware of alleged corruption by **X** who was arrested on charges of trying to sell Obama's senate seat."
- **Answer**: **Rod Blagojevich**



Code + Data: https://github.com/bdhingra/ga-reader

[Dhingra et al. 2017]

Gated Attention Readers

Table 1: Validation/Test accuracy (%) on WDW dataset for both "Strict" and "Relaxed" settings. Results with "†" are cf previously published works.

Model	**Strict**		**Relaxed**	
	Val	Test	Val	Test
Human †	–	84	–	–
Attentive Reader †	–	53	–	55
AS Reader †	–	57	–	59
Stanford AR †	–	64	–	65
NSE †	66.5	66.2	67.0	66.7
GA-- †	–	57	–	60.0
GA (update $L(w)$)	67.8	67.0	67.0	66.6
GA (fix $L(w)$)	68.3	68.0	69.6	69.1
GA (+feature, update $L(w)$)	70.1	69.5	70.9	71.0
GA (+feature, fix $L(w)$)	**71.6**	**71.2**	**72.6**	**72.6**

Table 2: **Top:** Performance of different gating functions. **Bottom:** Effect of varying the number of hops K. Results on WDW without using the qe-comm feature and with fixed $L(w)$.

Gating Function	**Accuracy**	
	Val	Test
Sum	64.9	64.5
Concatenate	64.4	63.7
Multiply	**68.3**	**68.0**
K		
1 (AS) †	–	57
2	65.6	65.6
3	**68.3**	68.0
4	**68.3**	**68.2**

[Dhingra et al. 2017]

Gated Attention Readers

Table 3: Validation/Test accuracy (%) on CNN, Daily Mail and CBT. Results marked with "†" are cf previously published works. Results marked with "‡" were obtained by training on a larger training set. Best performance on standard training sets is in bold, and on larger training sets in italics.

Model	**CNN**		**Daily Mail**		**CBT-NE**		**CBT-CN**	
	Val	Test	Val	Test	Val	Test	Val	Test
Humans (query) †	–	–	–	–	–	52.0	–	64.4
Humans (context + query) †	–	–	–	–	–	81.6	–	81.6
LSTMs (context + query) †	–	–	–	–	51.2	41.8	62.6	56.0
Deep LSTM Reader †	55.0	57.0	63.3	62.2	–	–	–	–
Attentive Reader †	61.6	63.0	70.5	69.0	–	–	–	–
Impatient Reader †	61.8	63.8	69.0	68.0	–	–	–	–
MemNets †	63.4	66.8	–	–	70.4	66.6	64.2	63.0
AS Reader †	68.6	69.5	75.0	73.9	73.8	68.6	68.8	63.4
DER Network †	71.3	72.9	–	–	–	–	–	–
Stanford AR (relabeling) †	73.8	73.6	77.6	76.6	–	–	–	–
Iterative Attentive Reader †	72.6	73.3	–	–	75.2	68.6	72.1	69.2
EpiReader †	73.4	74.0	–	–	75.3	69.7	71.5	67.4
AoA Reader †	73.1	74.4	–	–	77.8	72.0	72.2	69.4
ReasoNet †	72.9	74.7	77.6	76.6	–	–	–	–
NSE †	–	–	–	–	78.2	73.2	74.3	**71.9**
BiDAF †	76.3	76.9	80.3	79.6	–	–	–	–
MemNets (ensemble) †	66.2	69.4	–	–	–	–	–	–
AS Reader (ensemble) †	73.9	75.4	78.7	77.7	76.2	71.0	71.1	68.9
Stanford AR (relabeling,ensemble) †	77.2	77.6	80.2	79.2	–	–	–	–
Iterative Attentive Reader (ensemble) †	75.2	76.1	–	–	76.9	72.0	74.1	71.0
EpiReader (ensemble) †	–	–	–	–	76.6	71.8	73.6	70.6
AS Reader (+BookTest) † ‡	–	–	–	–	80.5	76.2	83.2	80.8
AS Reader (+BookTest,ensemble) † ‡	–	–	–	–	*82.3*	*78.4*	*85.7*	*83.7*
GA--	73.0	73.8	76.7	75.7	74.9	69.0	69.0	63.9
GA (update $L(w)$)	**77.9**	**77.9**	**81.5**	**80.9**	76.7	70.1	69.8	67.3
GA (fix $L(w)$)	77.9	77.8	80.4	79.6	77.2	71.4	71.6	68.0
GA (+feature, update $L(w)$)	77.3	76.9	80.7	80.0	77.2	73.3	73.0	69.8
GA (+feature, fix $L(w)$)	76.7	77.4	80.0	79.3	**78.5**	**74.9**	**74.4**	70.7

[Dhingra et al. 2017]

Adversarial Examples for Evaluating RC Systems

Article: Super Bowl 50
Paragraph: "*Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.*"
Question: "*What is the name of the quarterback who was 38 in Super Bowl XXXIII?*"
Original Prediction: John Elway
Prediction under adversary: Jeff Dean

[Jia and Liang, 2017]

Adversarial Examples for Evaluating RC Systems

	Image Classification	Reading Comprehension
Possible Input		Tesla moved to the city of Chicago in 1880.
Similar Input		Tadakatsu moved to the city of Chicago in 1881.
Semantics	Same	Different
Model's Mistake	Considers the two to be different	Considers the two to be the same
Model Weakness	Overly sensitive	Overly stable

Table 1: Adversarial examples in computer vision exploit model oversensitivity to small perturbations. In contrast, our adversarial examples work because models do not realize that a small perturbation can completely change the meaning of a sentence. Images from Szegedy et al. (2014).

[Jia and Liang, 2017]

Adversarial Examples for Evaluating RC Systems

Article: **Nikola Tesla**
Paragraph: "*In January 1880, two of Tesla's uncles put together enough money to help him leave Gospić for Prague where he was to study. Unfortunately, he arrived too late to enroll at Charles-Ferdinand University; he never studied Greek, a required subject; and he was illiterate in Czech, another required subject. Tesla did, however, attend lectures at the university, although, as an auditor, he did not receive grades for the courses.*"
Question: "*What city did Tesla move to in 1880?*"
Answer: *Prague*
Model Predicts: *Prague*

AddAny

Randomly initialize *d* words:

spring attention income getting reached

↓ Greedily change one word

spring attention income other reached

↓ Repeat many times

Adversary Adds: ***tesla move move other george***
Model Predicts: *george*

AddSent

What city did Tesla move to in 1880? — *Prague*

(Step 1) Mutate question

(Step 2) Generate fake answer

What city did Tadakatsu move to in 1881? — *Chicago*

(Step 3) Convert into statement

Tadakatsu moved the city of Chicago to in 1881.

(Step 4) Fix errors with crowdworkers, verify resulting sentences with other crowdworkers

Adversary Adds: ***Tadakatsu moved to the city of Chicago in 1881.***
Model Predicts: *Chicago*

[Jia and Liang, 2017]

Adversarial Examples for Evaluating RC Systems

Model	Original	ADDSENT	ADDONESENT
ReasoNet-E	**81.1**	39.4	49.8
SEDT-E	80.1	35.0	46.5
BiDAF-E	80.0	34.2	46.9
Mnemonic-E	79.1	**46.2**	**55.3**
Ruminating	78.8	37.4	47.7
jNet	78.6	37.9	47.0
Mnemonic-S	78.5	**46.6**	**56.0**
ReasoNet-S	78.2	39.4	50.3
MPCM-S	77.0	40.3	50.0
SEDT-S	76.9	33.9	44.8
RaSOR	76.2	39.5	49.5
BiDAF-S	75.5	34.3	45.7
Match-E	75.4	29.4	41.8
Match-S	71.4	27.3	39.0
DCR	69.3	37.8	45.1
Logistic	50.4	23.2	30.4

[Jia and Liang, 2017]

Multi-Hop QA

Paragraph A, Return to Olympus:
[1] *Return to Olympus is the only album by the alternative rock band Malfunkshun.* [2] *It was released after the band had broken up and after lead singer Andrew Wood (later of Mother Love Bone) had died of a drug overdose in 1990.* [3] Stone Gossard, of Pearl Jam, had compiled the songs and released the album on his label, Loosegroove Records.

Paragraph B, Mother Love Bone:
[4] *Mother Love Bone was an American rock band that formed in Seattle, Washington in 1987.* [5] The band was active from 1987 to 1990. [6] *Frontman Andrew Wood's personality and compositions helped to catapult the group to the top of the burgeoning late 1980s/early 1990s Seattle music scene.* [7] *Wood died only days before the scheduled release of the band's debut album, "Apple", thus ending the group's hopes of success.* [8] The album was finally released a few months later.

Q: What was the former band of the member of Mother Love Bone who died just before the release of "Apple"?
A: Malfunkshun
Supporting facts: 1, 2, 4, 6, 7

Figure 1: An example of the multi-hop questions in HotpotQA. We also highlight the supporting facts in *blue italics*, which are also part of the dataset.

[Yang et al., 2018]

Avoiding Reasoning Shortcuts in Multi-Hop QA

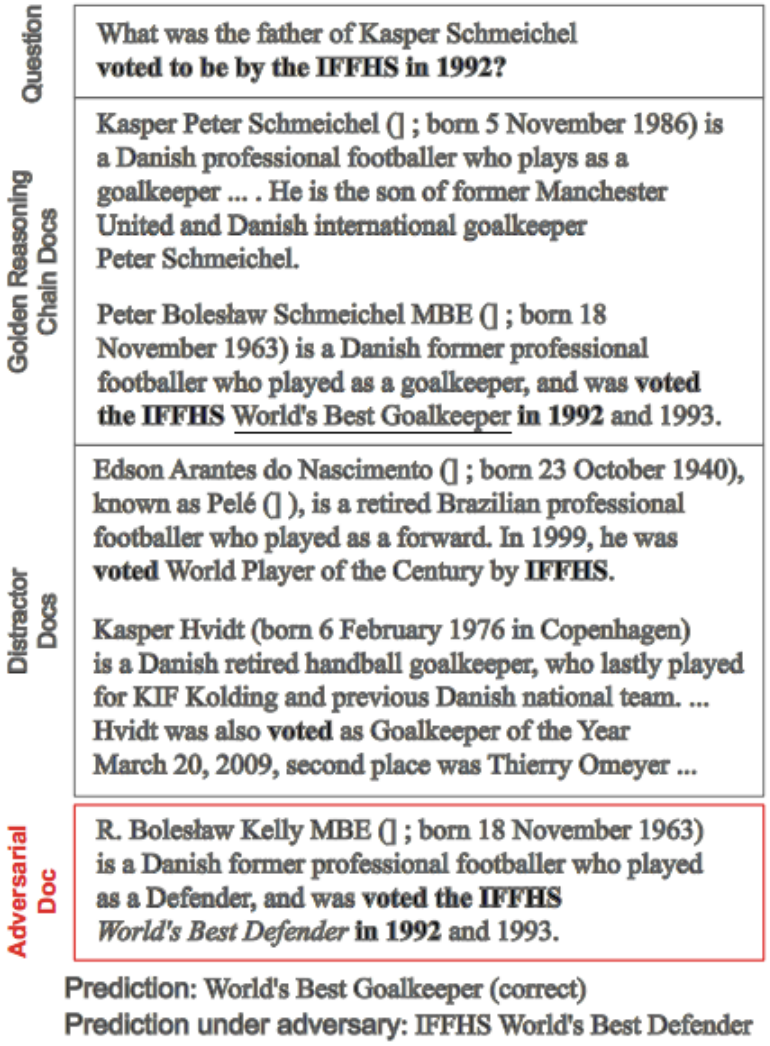


Figure 1: HotpotQA example with a reasoning shortcut, and our adversarial document that eliminates this shortcut to necessitate multi-hop reasoning.

[Chen & Bansal, 2019]

Neural Modular Networks for Multi-Hop QA

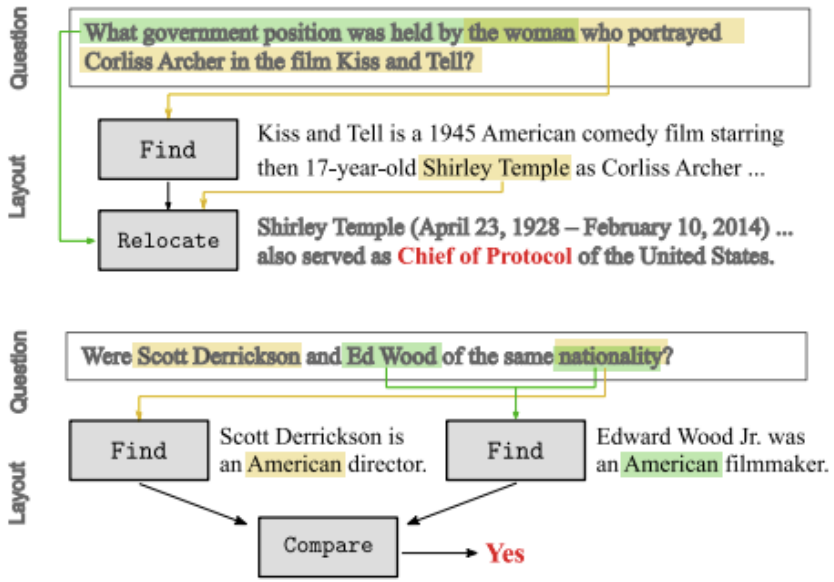


Figure 1: Two HotpotQA examples and the modular network layout predicted by the controller.

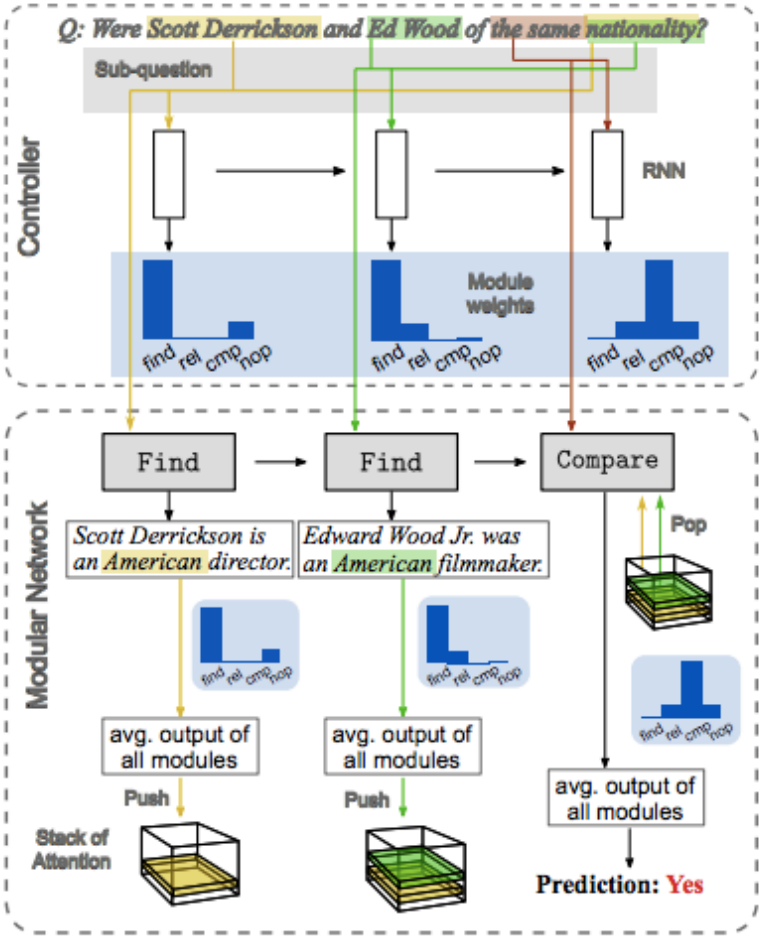


Figure 2: Modular network with a controller (top) and the dynamically-assembled modular network (bottom). At every step, the controller produces a sub-question vector and predicts a distribution to weigh the averages of the modules' outputs.

[Chen & Bansal, 2019b]

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 (here we use our semantic parsers discussed previously)

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]