Question Answering

Zhiyuan Liu
liuzy@tsinghua.edu.cn
THUNLP

Outline

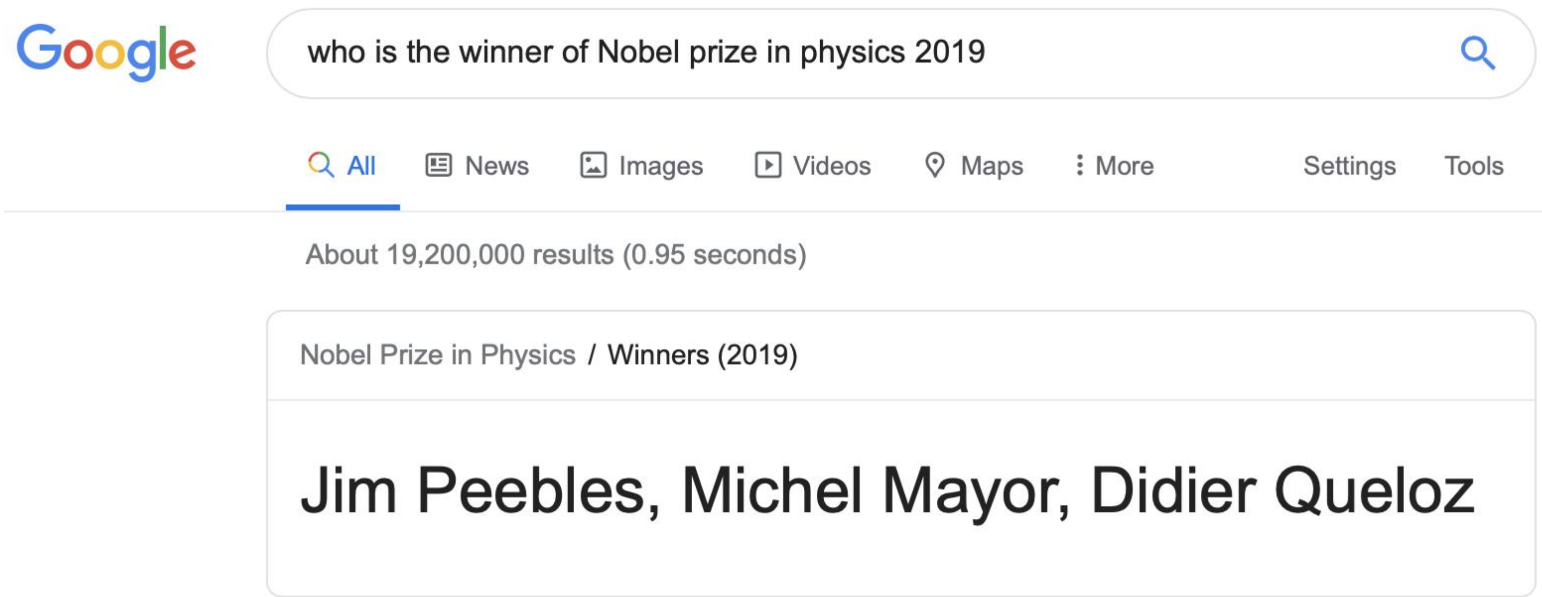
- Introduction to QA
- Reading Comprehension
- Open-domain QA
- KBQA

Background

- Why we need question answering (QA) ?
  - When we search for something in Google, it's usually hard to find answers from the document list.
  - With QA systems, answers are automatically found from large amount of data

Applications of QA

- Better search experience



Applications of QA

- IBM Watson: 2011 Winner in Jeopardy
- Defeat two human players (Ken and Brad)

Applications of QA

- Intelligent assistants

Apple Siri (2011)

Microsoft Cortana (2014)

Amazon Alexa (2014)

Google Home (2016)

History

Template-based QA Expert System

1960

IR-based QA

1990

Community QA

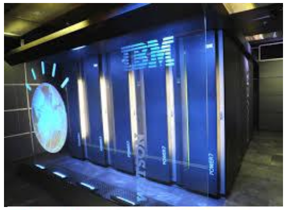
2000

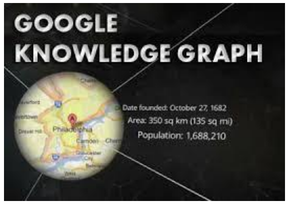
2010

Machine Reading Comprehension KBQA

BASEBALL
LUNAR
MACSYMA
SHRDLE

MASQUE
TREC





Types of QA

- Machine Reading Comprehension:
  - Read specific documents and answer questions.
- Open-domain QA:
  - Search and read relevant documents to answer questions.
- Knowledge-based QA:
  - Answer questions based on knowledge graph.
- Conversational QA and dialog:
  - Answer questions according to dialog history.
- …

Outline

- Introduction to QA
- Reading Comprehension
  - Task Definition, Dataset and Evaluation
  - Model Framework
  - Multi-paragraph
- Open-domain QA
- KBQA

Definition of RC

Documents

One night I was at my friend's house where he threw a party. We were enjoying our dinner at night when all of a sudden, we heard a knock on the door. I opened the door and saw this guy who had scar on his face. As soon as I saw him, I ran inside the house and called the cops. The cops came and the guy ran away as soon as he heard the cop car coming. We never found out what happened to that guy after that day.

Questions

Candidate answers

1. What was the strange guy doing with the friend?
A) enjoying a meal
B) talking about his job
C) talking to him
D) trying to beat him

2. Why did the strange guy run away?
A) because he heard the cop car
B) because he saw his friend
C) because he didn't like the dinner
D) because it was getting late

Types of RC

- Cloze test
  - CNN/Daily Mail (93k CNN articles, 220k Daily Mail articles)

	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*

Types of RC

- Cloze test
  - CBT (Children's Book Test)
  - Context: 20 continuous sentences
  - Question: the 21st sentence, with an entity masked
  - Answer: the masked entity
  - 10 candidates

S: 1 Mr. Cropper was opposed to our hiring you .
2 Not , of course , that he had any personal objection to you , but he is set against female teachers , and when a Cropper is set there is nothing on earth can change him .
3 He says female teachers ca n't keep order .
4 He 's started in with a spite at you on general principles , and the boys know it .
5 They know he 'll back them up in secret , no matter what they do , just to prove his opinions .
6 Cropper is sly and slippery , and it is hard to corner him . ''
7 `` Are the boys big ? ''
8 queried Esther anxiously .
9 `` Yes .
10 Thirteen and fourteen and big for their age .
11 You ca n't whip 'em -- that is the trouble .
12 A man might , but they 'd twist you around their fingers .
13 You 'll have your hands full , I 'm afraid .
14 But maybe they 'll behave all right after all . ''
15 Mr. Baxter privately had no hope that they would , but Esther hoped for the best.
16 She could not believe that Mr. Cropper would carry his prejudices into a personal application .
17 This conviction was strengthened when he overtook her walking from school the next day and drove her home .
18 He was a big , handsome man with a very suave , polite manner .
19 He asked interestedly about her school and her work , hoped she was getting on well , and said he had two young rascals of his own to send soon .
20 Esther felt relieved .

q: She thought that Mr. ________ had exaggerated matters a little .

C: Baxter, Cropper, Esther, course, fingers, manner, objection, opinion, right, spite.

a: Baxter

Types of RC

- Multiple choice
  - RACE: 100k multiple choice questions collected from English exams in China.

Passage:

In a small village in England about 150 years ago, a mail coach was standing on the street. It didn't come to that village often. People had to pay a lot to get a letter. The person who sent the letter didn't have to pay the postage, while the receiver had to.

"Here's a letter for Miss Alice Brown," said the mailman.

" I'm Alice Brown," a girl of about 18 said in a low voice.

Alice looked at the envelope for a minute, and then handed it back to the mailman.

"I'm sorry I can't take it, I don't have enough money to pay it", she said.

A gentleman standing around were very sorry for her. Then he came up and paid the postage for her.

When the gentleman gave the letter to her, she said with a smile, " Thank you very much, This letter is from Tom. I'm going to marry him. He went to London to look for work. I've waited a long time for this letter, but now I don't need it, there is nothing in it."

"Really? How do you know that?" the gentleman said in surprise.

"He told me that he would put some signs on the envelope. Look, sir, this cross in the corner means that he is well and this circle means he has found work. That's good news."

The gentleman was Sir Rowland Hill. He didn't forgot Alice and her letter.

"The postage to be paid by the receiver has to be changed," he said to himself and had a good plan.

"The postage has to be much lower, what about a penny? And the person who sends the letter pays the postage. He has to buy a stamp and put it on the envelope." he said . The government accepted his plan. Then the first stamp was put out in 1840. It was called the "Penny Black". It had a picture of the Queen on it.

Questions:

1): The first postage stamp was made _.
A. in England B. in America C. by Alice D. in 1910

2): The girl handed the letter back to the mailman because _ .
A. she didn't know whose letter it was
B. she had no money to pay the postage
C. she received the letter but she didn't want to open it
D. she had already known what was written in the letter

3): We can know from Alice's words that _ .
A. Tom had told her what the signs meant before leaving
B. Alice was clever and could guess the meaning of the signs
C. Alice had put the signs on the envelope herself
D. Tom had put the signs as Alice had told him to

4): The idea of using stamps was thought of by _ .
A. the government
B. Sir Rowland Hill
C. Alice Brown
D. Tom

5): From the passage we know the high postage made _ .
A. people never send each other letters
B. lovers almost lose every touch with each other
C. people try their best to avoid paying it
D. receivers refuse to pay the coming letters

Answer: ADABC

Types of RC

- Extractive RC: Predict a span in documents
  - SQuAD: 10k human-annotated questions and 536 articles from Wikipedia. Every answer is a span in the article

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? Answer: gravity

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

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

Evaluation of Extractive RC

- EM (Exact Match): The percentage of predictions that match any one of the ground truth answers exactly.
- F1: The average overlap between the prediction and ground truth answer.
- Examples:

Ground truth	Prediction	EM	F1
France	France	1	1
10th, 10th century	10th	1	1
Richard I	Richard I of Normandy	0	0.67
the Pechenegs	Pechenegs	1	1

Datasets

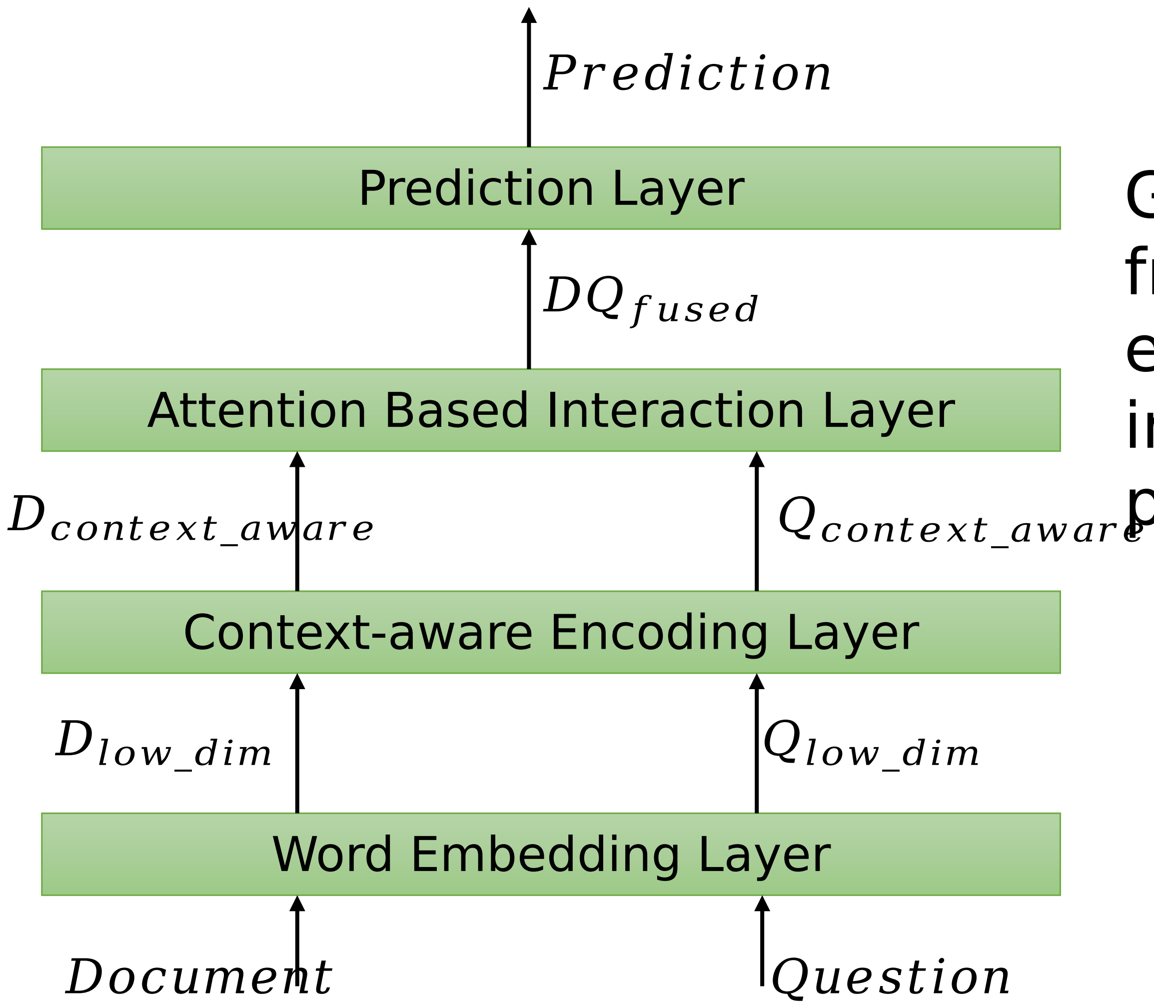
Multiple choice
Cloze test
Extractive

2013	MCTest
2015	bAbI, CNN/Daily Mail
2016	SCT, LAMBDA, MSMARCO, NEWSQA, SQuAD, CBT
2017	Who-Did-What, SearchQA, NarrativeQA, DuReader, RACE, Quasar, TriviaQA
2018	CoQA, HOTPOTQA, SQuAD2.0, CliCR, DuoRC, CLOTH, MCScript, ARC, emrQA, RecipeQA, openBookQA, ShARC, QuAC, ProPara, TextWordsQA
2019	DREAM, DROP, Natural Questions, RC-QED, GeosQA, CosmosQA, QASC

Outline

- Introduction to QA
- Reading Comprehension
  - Task Definition, Dataset and Evaluation
  - Model Framework
  - Multi-paragraph
- Open-domain QA
- KBQA

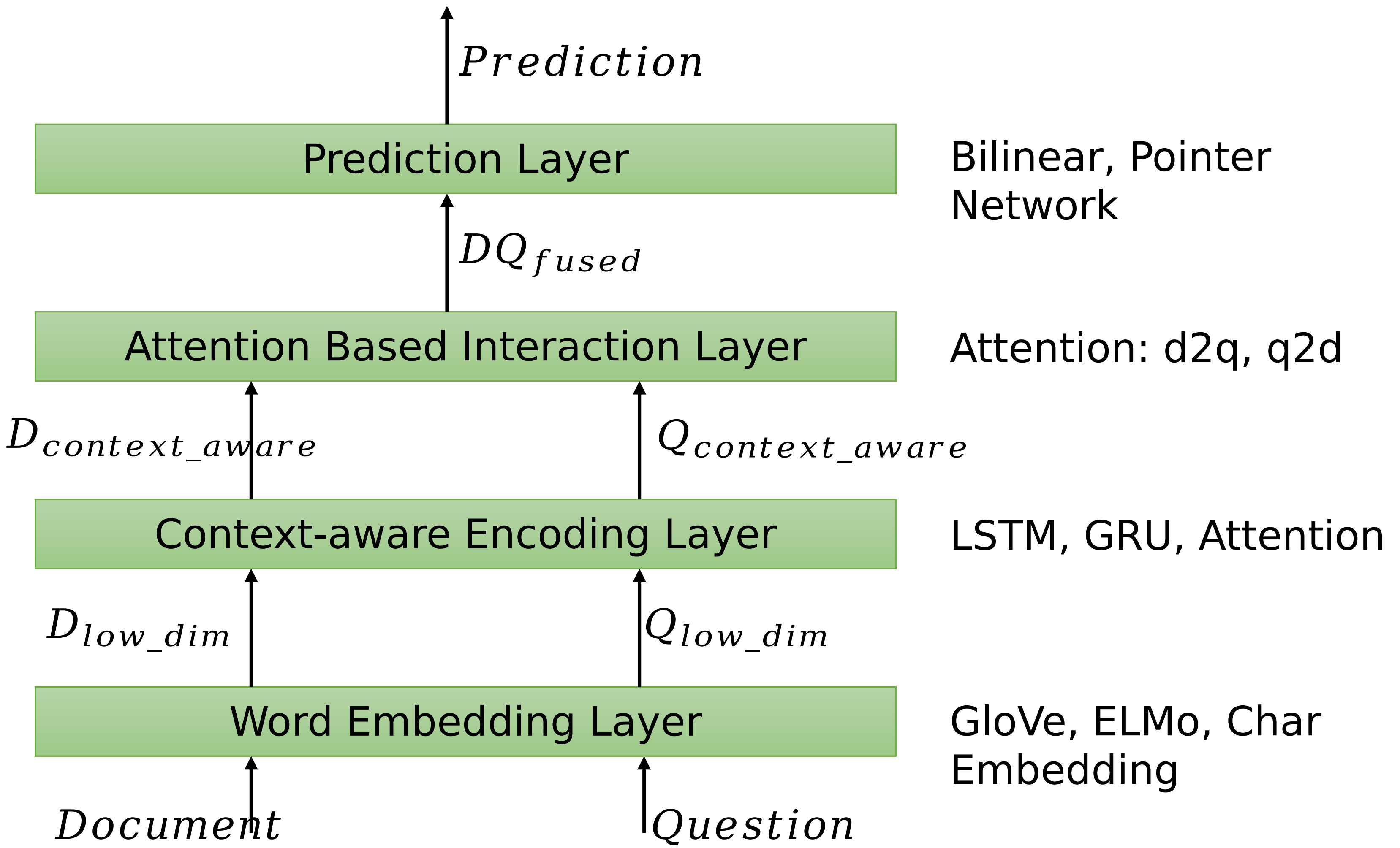
Model Framework



General framework in RC: embed, encode, interact, and predict

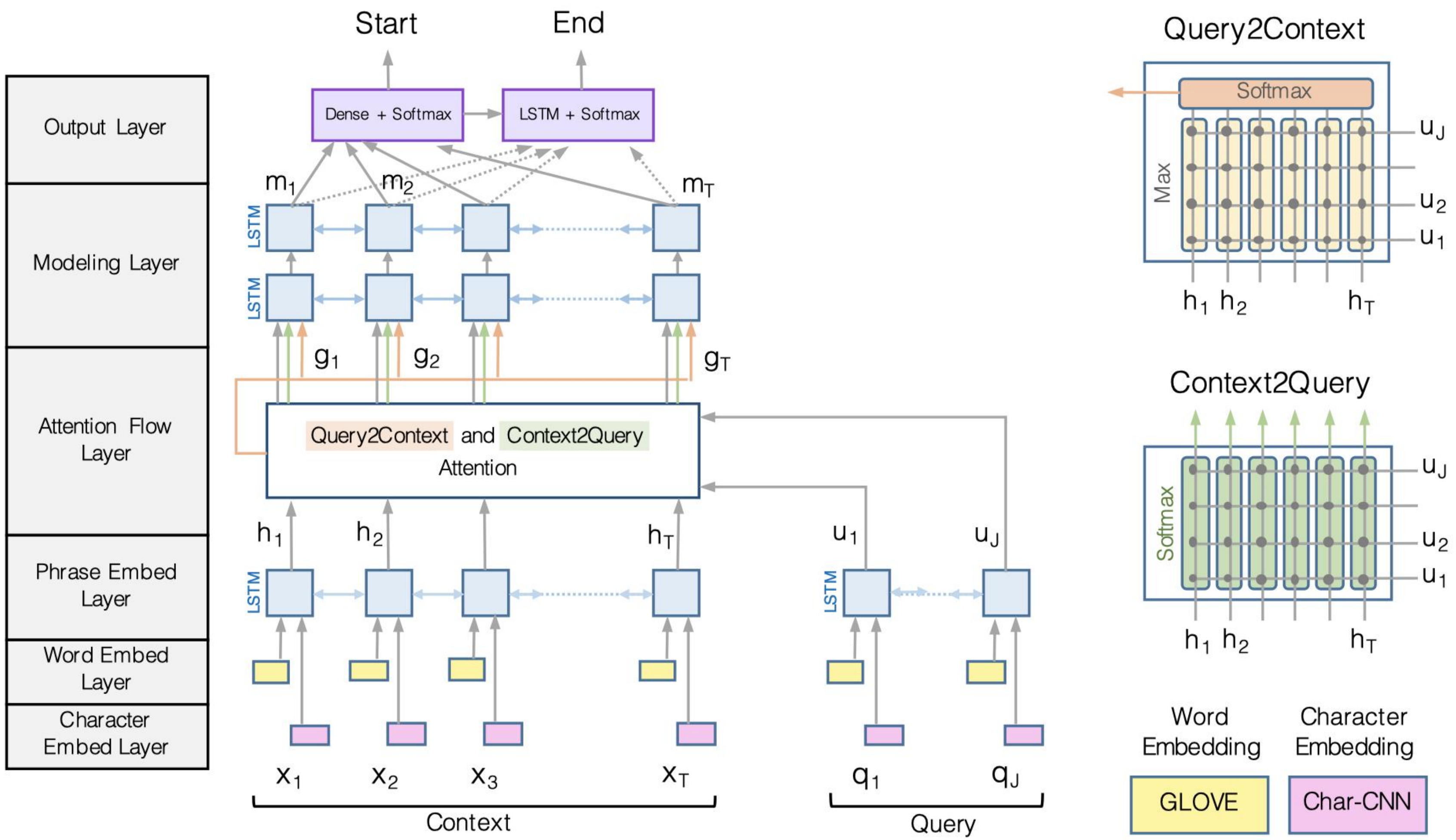
Seo et al., Bidirectional Attention Flow for Machine Comprehension, In Proceedings of ICLR 2017

Model Framework



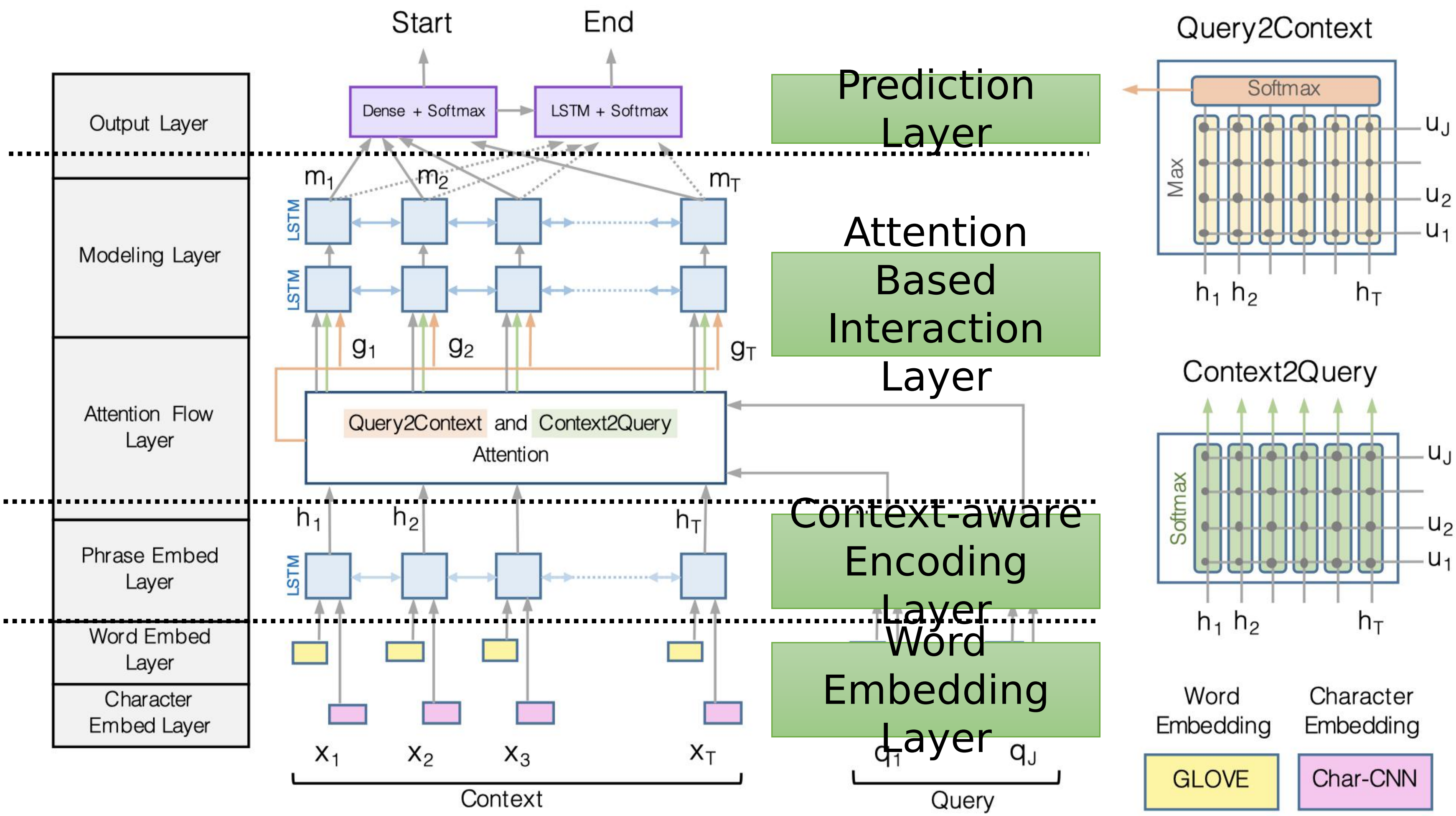
Seo et al., Bidirectional Attention Flow for Machine Comprehension, In Proceedings of ICLR 2017

An Example of RC Model: BiDAF



Seo et al., Bidirectional Attention Flow for Machine Comprehension, In Proceedings of ICLR 2017

An Example of RC Model: BiDAF



Seo et al., Bidirectional Attention Flow for Machine Comprehension, In Proceedings of ICLR 2017

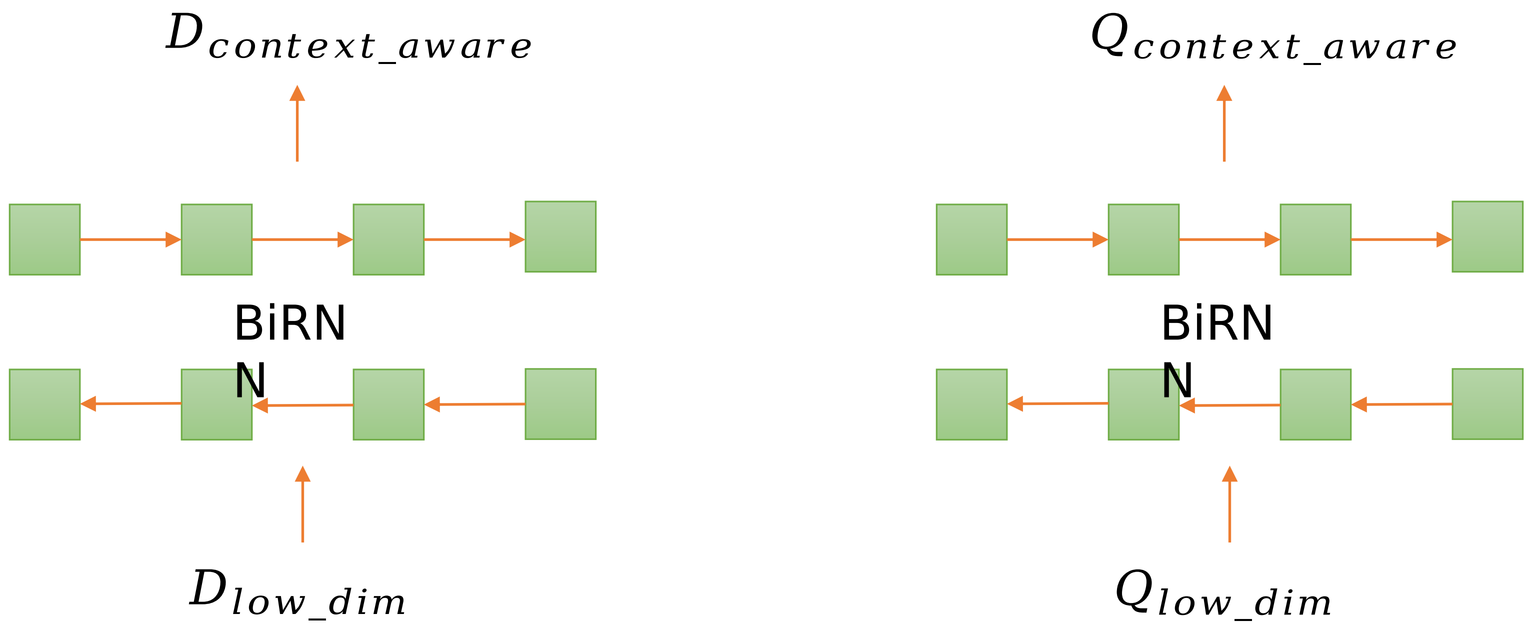
Word Embedding Layer

- How can we get the word representation?
- Word Embedding (Glove, word2vec)
- Word Embedding + Character Embedding
  - Obtain the character embedding of each word using CNN + Max-Pooling.
  - The word embedding and character embedding are passed to Highway Network.
- Pre-trained Word Embedding (ELMo)

Seo et al., Bidirectional Attention Flow for Machine Comprehension, In Proceedings of ICLR 2017

Context-aware Encoding Layer

- Bidirectional RNN are widely used in contextual encoding



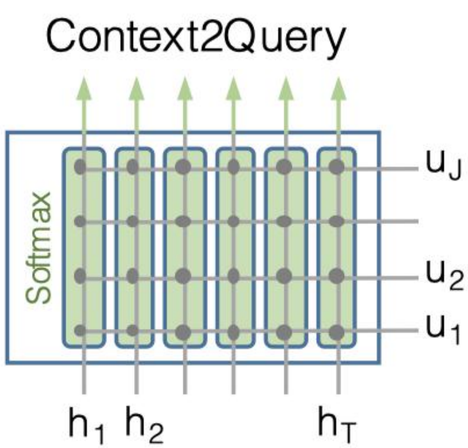
Seo et al., Bidirectional Attention Flow for Machine Comprehension, In Proceedings of ICLR 2017

Attention Based Interaction Layer: Q-C Attention

- The inputs are context-aware representations of context $H \in R^{T \times d}$ and question $U \in R^{J \times d}$.
- Compute a similarity matrix $S \in R^{T \times J}$:

$$S_{tj} = \alpha(H_t, U_j) = w_s^T [H_t; U_j; H_t \circ U_j]$$

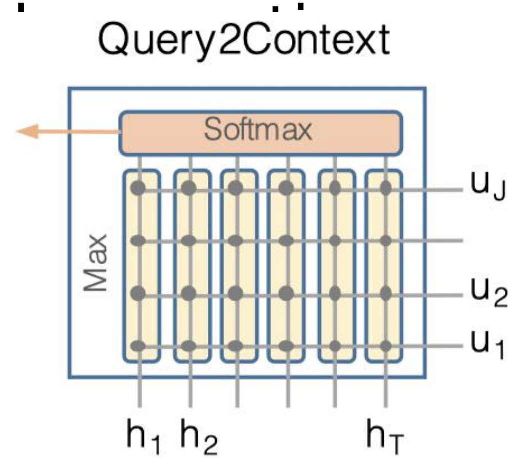
- Context2Question Attention:
  - Signify which question word is most relevant to each context word.
  - $a_t = \mathrm{softmax}(S_{t:})$
  - $\widetilde{U}_t = \sum_j a_{tj} U_j \in R^{T \times d}$



Seo et al., Bidirectional Attention Flow for Machine Comprehension, In Proceedings of ICLR 2017

Attention Based Interaction Layer: Q-C Attention

- Question2Context Attention:
  - Find which context words has the closest similarity to one of the question words and are hence critical for answering
  - $b = \text{softmax}\left(\max_{col} S\right) \in R^T$
  - $\tilde{h} = \sum_t b_t h_t \in R^d$

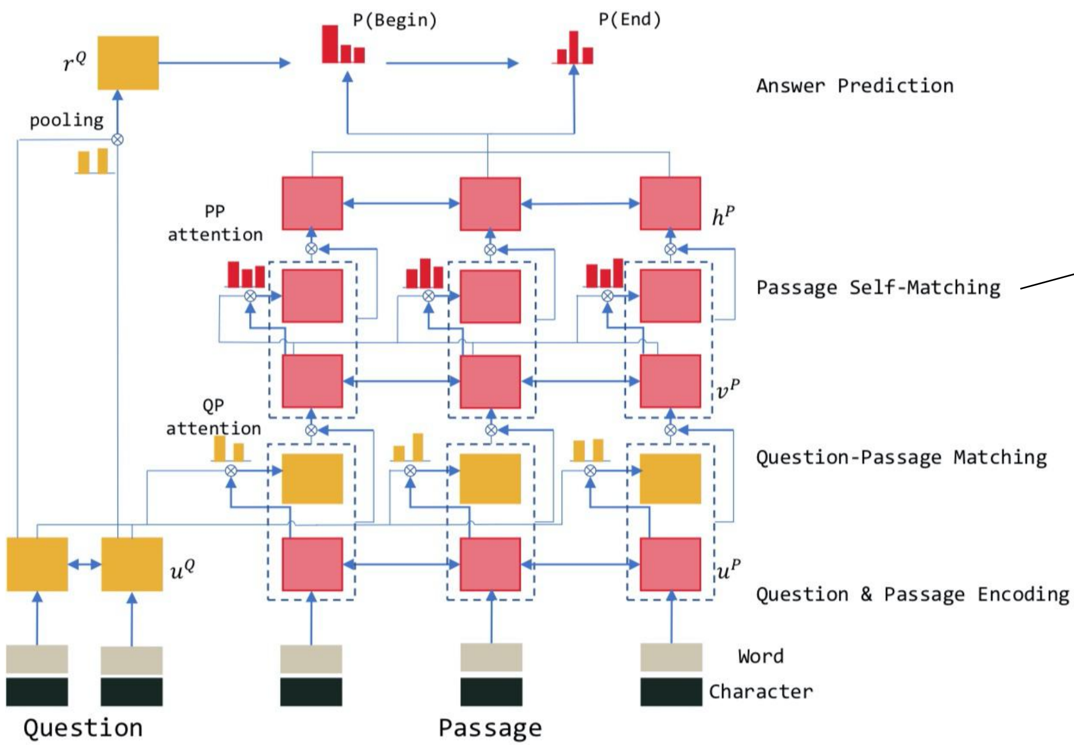


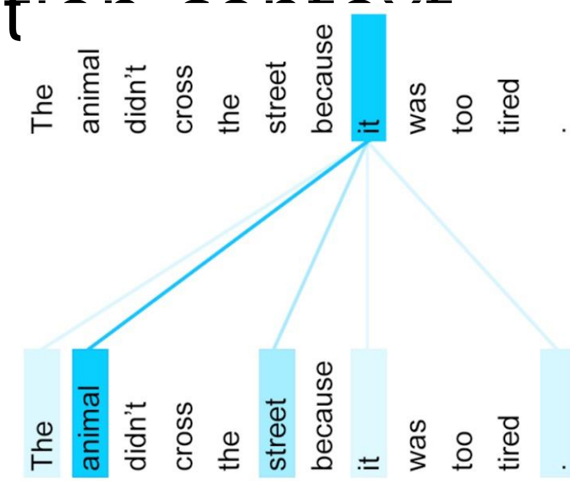
- Question-aware representa
  - $G_t = [H_t; \widetilde{U}_t; H_t \circ \widetilde{U}_t; H_t \circ \hat{h}$
  - $M_1, \ldots, M_T = \text{BiRNN}(G_1, \ldots, G_T)$
  - M_i is expected to contain contextual information about the word with respect to the entire context and the question

Seo et al., Bidirectional Attention Flow for Machine Comprehension, In Proceedings of ICLR 2017

Attention Based Interaction Layer: Self-Attention

- Self-attention: capture contextual information
- Apply self-attention after question context





$$s_{tj} = v^T \tanh\left(W_v v_j^P + \widetilde{W}_v v_t^P\right)$$

$$a_{t:} = \mathrm{softmax}(s_{t:})$$

$$c_t = \Sigma_i a_{ti} v_i^P$$

$$h_t^P = \mathrm{BiRNN}([\ldots, \ldots^P])$$

Wang, et al., Gated Self-Matching Networks for Reading Comprehension and Question Answering, In Proceedings of ACL 2017

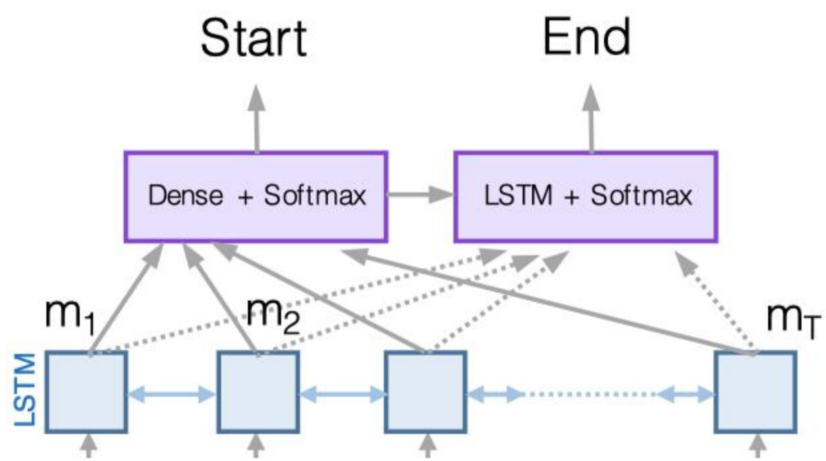
Prediction Layer

- Extractive RC
  - Use pointer network to find the start and end of the answer span.
  - We obtain the probability distribution of the start index over the entire paragraph by

$$p^1 = \mathrm{softmax}(w_{p_1}^T [G; M])$$

  - We get p^2 in a similar way, with an additional BiRNN on M.
  - Loss function:

$$L = -\frac{1}{N}\sum_i \log p^1_{y_i^1} + \log p^2_{y_i^2}$$



Prediction Layer

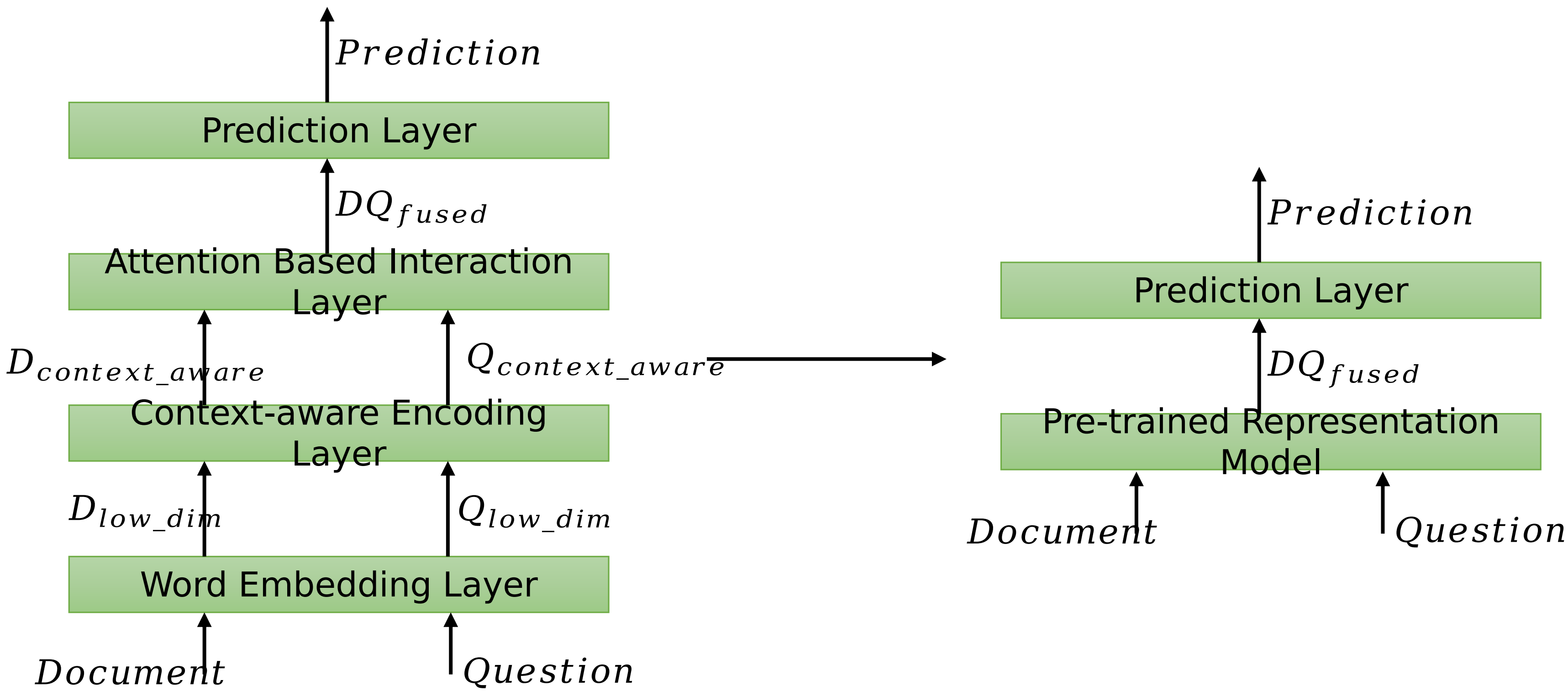
- Cloze Test
  - Since each answer in the CNN/DailyMail and CBT is always a single word (entity) in paragraph, we only need to predict the start index.
  - We mask out all words that are not candidates.

	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*

Seo et al., Bidirectional Attention Flow for Machine Comprehension, In Proceedings of ICLR 2017

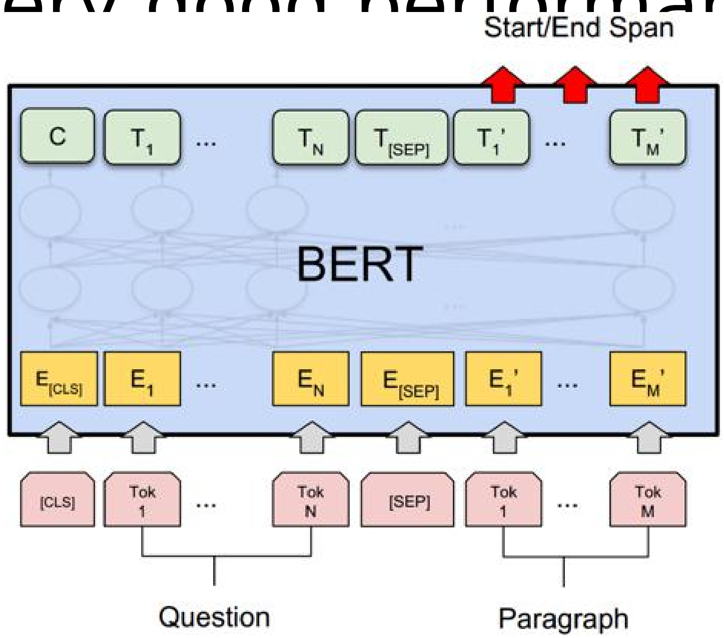
Attention is All You Need

- Use PLMs (like BERT) to replace the first three layers
  - The BERT-based model has no RNN modules



BERT-based Model

- Feed the concatenation of the question and the context to BERT. Get the question-aware context representation to predict the start/end of answers.
- Very good performance



Leaderboard

SQuAD2.0 tests the ability of a system to not only answer reading comprehension questions, but also abstain when presented with a question that cannot be answered based on the provided paragraph. How will your system compare to humans on this task?

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Nov 08, 2018	BERT (single model) *Google AI Language*	**80.005**	**83.061**
2 Nov 06, 2018	SLQA+BERT (single model) *Alibaba DAMO NLP* http://www.aclweb.org/anthology/P18-1158	77.003	80.209
3 Nov 08, 2018	BERT_base_aug (ensemble) *GammaLab*	76.721	79.611
4 Nov 05, 2018	MIR-MRC(F-Net) (single model) *Kangwon National University, Natural Language Processing Lab. & ForceWin, KP Lab.*	74.803	77.988
5 Sep 13, 2018	nlnet (single model) *Microsoft Research Asia*	74.238	77.022

Relevance rather than Understanding

- Adding adversarial examples to the original documents would impair system performance seriously.

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

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. Adversarial Examples for Evaluating Reading Comprehension Systems. In Proceedings of EMNLP 2017

Need Deep Reasoning

- Reasoning from implicit commonsense knowledge
- Answering beyond lexical matching
- Inference from multiple sentences
- Employing external knowledge
- Sentiment, Event, etc.
- Discourse relation, co-reference, etc.

Outline

- Introduction to QA
- Reading Comprehension
  - Task Definition, Dataset and Evaluation
  - Model Framework
  - Multi-paragraph
- Open-domain QA
- KBQA

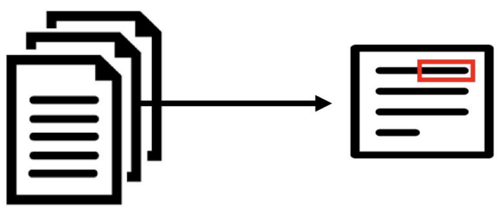
Challenge: Scaling Models to Documents

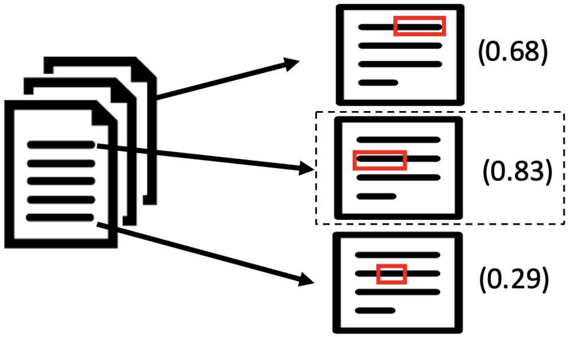
- Reading comprehension models have many layers and parameters
- Efficiency reduces as the paragraph length increases due to long RNN chains or transformers/self-attention modules
- Limits the model to processing short paragraphs

Clark and Gardner. Simple and Effective Multi-Paragraph Reading Comprehension. In Proceedings of ACL 2018

Challenge: Scaling Models to Documents

- Possible Approaches
  - Pipelined Systems
    - Select a single paragraph from the input, and run the model on that paragraph
  - Confidence Systems
    - Run the model on many paragraphs from the input, and make it assign a confidence score to its results on each paragraph





Clark and Gardner. Simple and Effective Multi-Paragraph Reading Comprehension. In Proceedings of ACL 2018

Pipelined Systems

- Paragraph selection
  - Train a shallow linear model to select the best paragraphs
  - Features include TF-IDF, word occurrences, and its position within the document
- Noisy Supervision
  - We label all text spans th
    as being correct
  - Training objective
    - $-\log \sum_{s \in \mathrm{S}} p(s)$

Question: Which British general was killed at Khartoum in 1885?
Answer: Gordon
Context: In February 1885 Gordon returned to the Sudan to evacuate Egyptian forces. Khartoum came under siege the next month and rebels broke into the city, killing Gordon and the other defenders. The British public reacted to his death by acclaiming 'Gordon of Khartoum', a saint. However, historians have suggested that Gordon...

Confidence Methods

- We can derive confidence scores from the logit scores given to each span by the model
- However, a model can be less confident in a correct extraction from one paragraph than in an incorrect extraction from another paragraph

Question	Low Confidence Correct Extraction	High Confidence Incorrect Extraction
When is the Members Debate held?	**Immediately after Decision Time** a "Members Debate" is held, which lasts for 45 minutes...	...majority of the Scottish electorate voted for it in a referendum to be held on **1 March 1979** that represented at least...
How many tree species are in the rainforest?	...one 2001 study finding a quarter square kilometer (62 acres) of Ecuadorian rainforest supports more than **1,100** tree species	The affected region was approximately **1,160,000** square miles (3,000,000 km2) of rainforest, compared to 734,000 square miles
Who was Warsz?	....In actuality, Warsz was a 12th/13th century **nobleman** who owned a village located at the modern....	One of the most famous people born in Warsaw was **Maria Sklodowska - Curie**, who achieved international...
How much did the initial LM weight in kg?	The initial LM model weighed approximately 33,300 pounds (**15,000** kg), and...	The module was 11.42 feet (3.48 m) tall, and weighed approximately 12,250 pounds (**5,560** kg)

Clark and Gardner. Simple and Effective Multi-Paragraph Reading Comprehension. In Proceedings of ACL 2018

Learning Well-Calibrated Confidence Scores

- Train the model on both answering-containing and non-answering containing paragraph and use a modified objective function
  - **Merge**: Concatenate sampled paragraphs together
  - **Non-Answer**: Process paragraphs independently, and allow the model to place probability mass on a “non-answer” output
  - **Sigmoid**: Assign an independent probability on each span using the sigmoid operator
  - **Shared-Norm**: Process paragraphs independently, but compute the span probability across spans in all paragraphs

Clark and Gardner. Simple and Effective Multi-Paragraph Reading Comprehension. In Proceedings of ACL 2018.

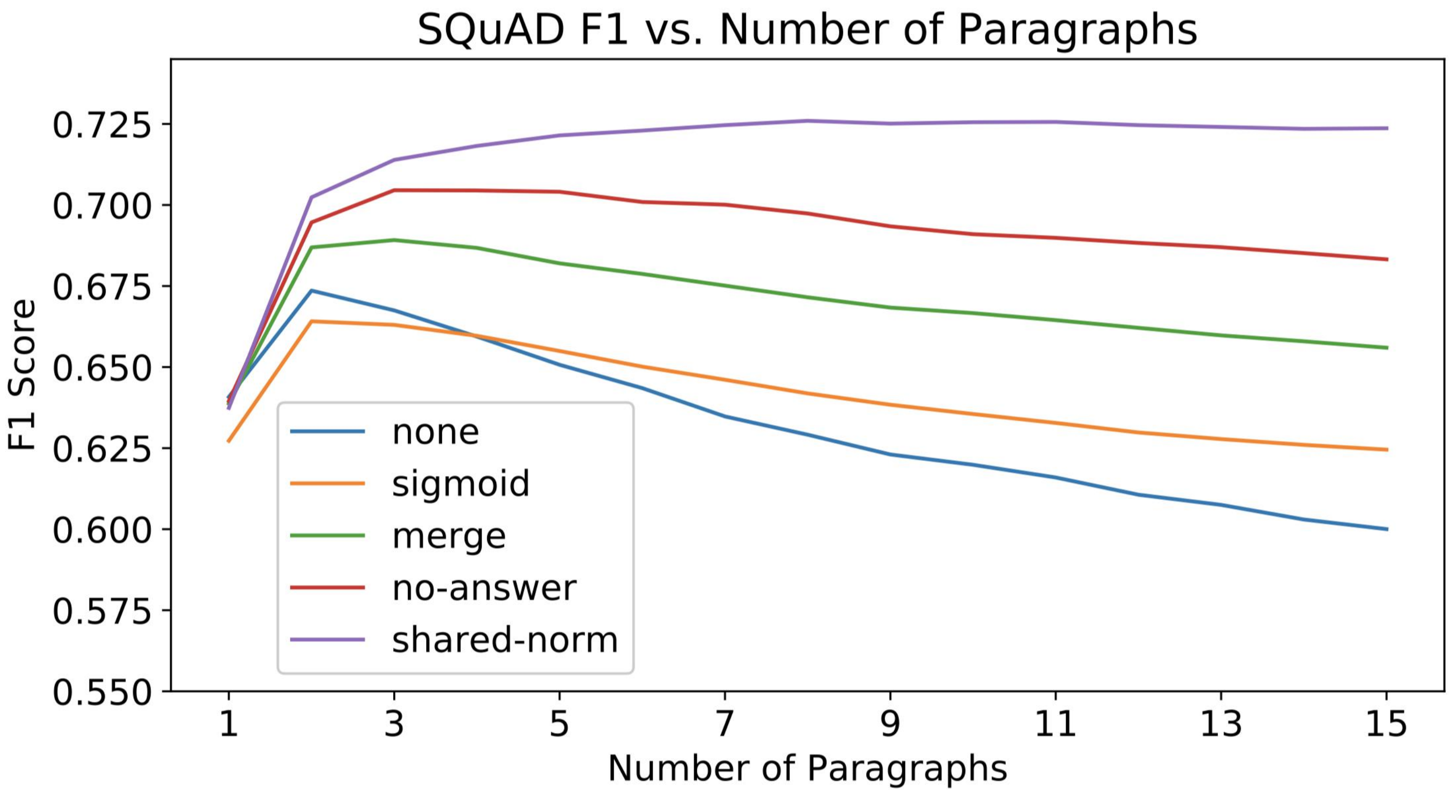
Datasets

- TriviaQA: Dataset of trivia questions and related documents found by web-search, including 3 settings
  - Web (a single document for each questions)
  - Wiki (multiple wikipedia documents for each questions)
  - Unfiltered (multiple documents for each questions)
- SQuAD: Turker-generated questions about Wikipedia articles
  - We use the questions paired with the entire article
  - Manual annotation shows most (90%) of questions are answerable as given the document it was generated from

Clark and Gardner. Simple and Effective Multi-Paragraph Reading Comprehension. In Proceedings of ACL 2018

Results

- Shared-Norm has best results



Clark and Gardner. Simple and Effective Multi-Paragraph Reading Comprehension. In Proceedings of ACL 2018.

Outline

- Introduction to QA
- Reading Comprehension
- Open-domain QA
- KBQA

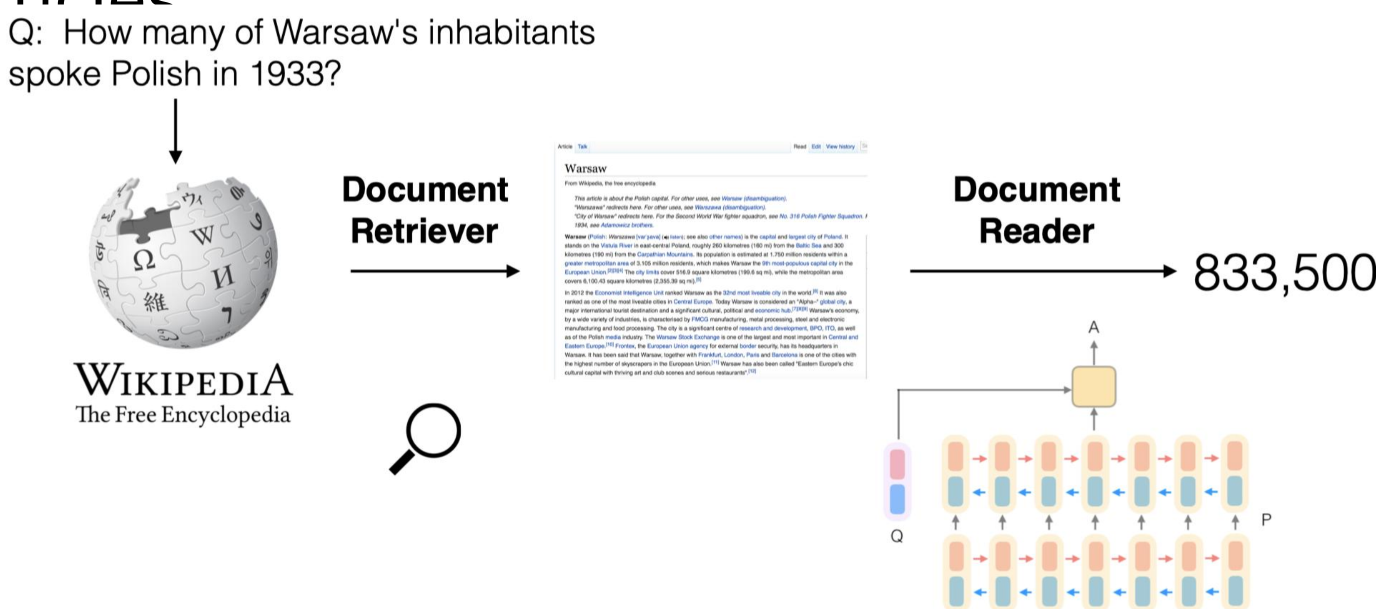
Open-domain QA

- RC assume that any question has a short piece of relevant text, which is impractical
- In open-domain QA, the model should be able to find relevant texts from a corpus and read them
  - Wikipedia can be viewed as a large-scale corpus for factoid question
- Goal: build an end-to-end QA system that can use full Wikipedia to answer any factoid question

Chen et al., Reading Wikipedia to Answer Open-Domain Questions. In Proceedings of ACL 2017

Open-domain QA

- Document Retriever + Document Reader
  - Document retriever: finding relevant articles from 5 million Wikipedia articles
  - Document reader (reading comprehension system): identifying the answer spans from those articles



Chen et al., Reading Wikipedia to Answer Open-Domain Questions. In Proceedings of ACL 2017

Document Retriever

- Return 5 Wikipedia articles given any question
- Features:
  - TF-IDF bag-of-words vectors
  - Efficient bigram hashing (Weinberger et al., 2009)
- Better performance than Wikipedia search: (hit@5)

Dataset	Wiki Search	Doc. Retriever plain	Doc. Retriever +bigrams
SQuAD	62.7	76.1	**77.8**
CuratedTREC	81.0	85.2	**86.0**
WebQuestions	73.7	**75.5**	74.4
WikiMovies	61.7	54.4	**70.3**

Chen et al., Reading Wikipedia to Answer Open-Domain Questions. In Proceedings of ACL 2017

Document Reader

- Simple reading comprehension model
- Features:
  - Word embeddings
  - Exact match features: whether the word appears in question
  - Token features: POS, NER, term frequency
  - Aligned question embedding
- Using Shared-Norm for multi-documents

Chen et al., Reading Wikipedia to Answer Open-Domain Questions. In Proceedings of ACL 2017

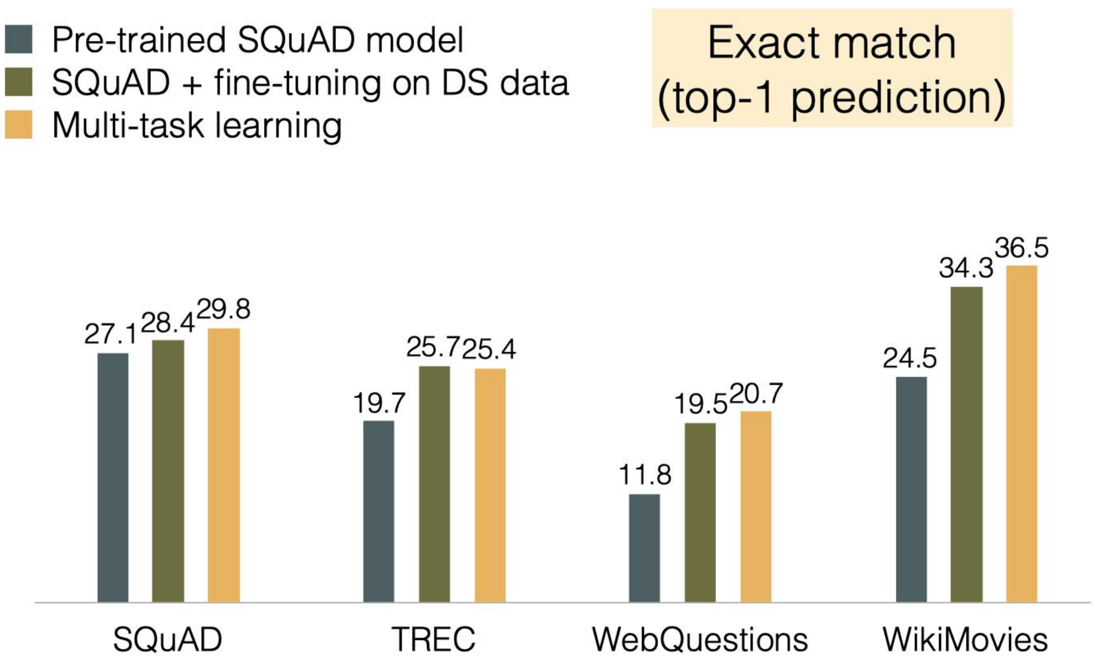
Distance Supervision

- For given question, automatically associate paragraphs including the answer span to this q

Dataset	Example	Article / Paragraph
SQuAD	**Q**: How many provinces did the Ottoman empire contain in the 17th century? **A**: 32	**Article**: Ottoman Empire **Paragraph**: ... At the beginning of the 17th century the empire contained 32 provinces and numerous vassal states. Some of these were later absorbed into the Ottoman Empire, while others were granted various types of autonomy during the course of centuries.
CuratedTREC	**Q**: What U.S. state's motto is "Live free or Die"? **A**: New Hampshire	**Article**: Live Free or Die **Paragraph**: "Live Free or Die" is the official motto of the U.S. state of New Hampshire, adopted by the state in 1945. It is possibly the best-known of all state mottos, partly because it conveys an assertive independence historically found in American political philosophy and partly because of its contrast to the milder sentiments found in other state mottos.
WebQuestions	**Q**: What part of the atom did Chadwick discover?[†] **A**: neutron	**Article**: Atom **Paragraph**: ... The atomic mass of these isotopes varied by integer amounts, called the whole number rule. The explanation for these different isotopes awaited the discovery of the neutron, an uncharged particle with a mass similar to the proton, by the physicist James Chadwick in 1932. ...
WikiMovies	**Q**: Who wrote the film Gigli? **A**: Martin Brest	**Article**: Gigli **Paragraph**: Gigli is a 2003 American romantic comedy film written and directed by Martin Brest and starring Ben Affleck, Jennifer Lopez, Justin Bartha, Al Pacino, Christopher Walken, and Lainie Kazan.

Results

- Reasonable performance across all four datasets
- Models using DS outperform models trained on SQuAD
  - Multi-



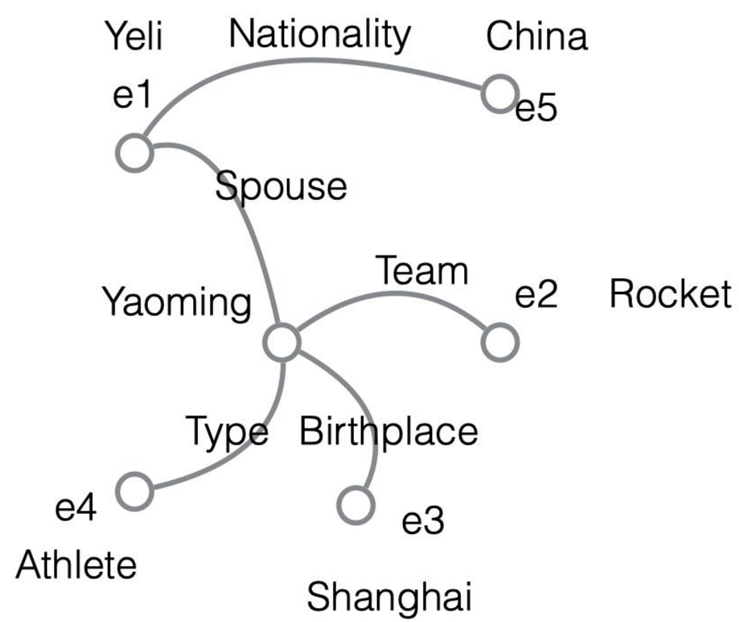
Chen et al., Reading Wikipedia to Answer Open-Domain Questions. In Proceedings of ACL 2017

Outline

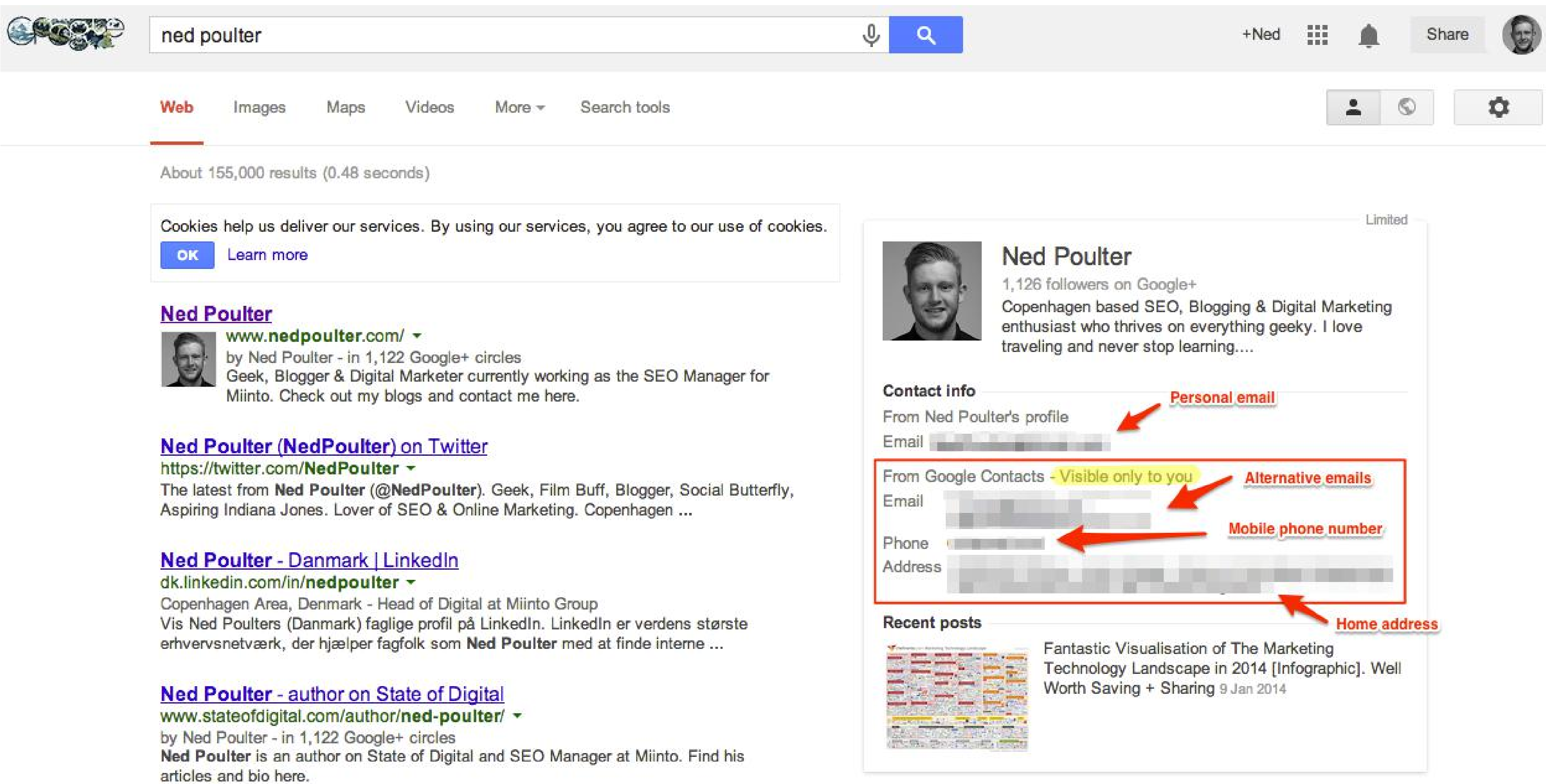
- Introduction to QA
- Reading Comprehension
- Open-domain QA
- KBQA
  - Definition and Dataset
  - Semantic Parsing for KBQA
  - Deep Learning for KBQA

KB-based QA: Example

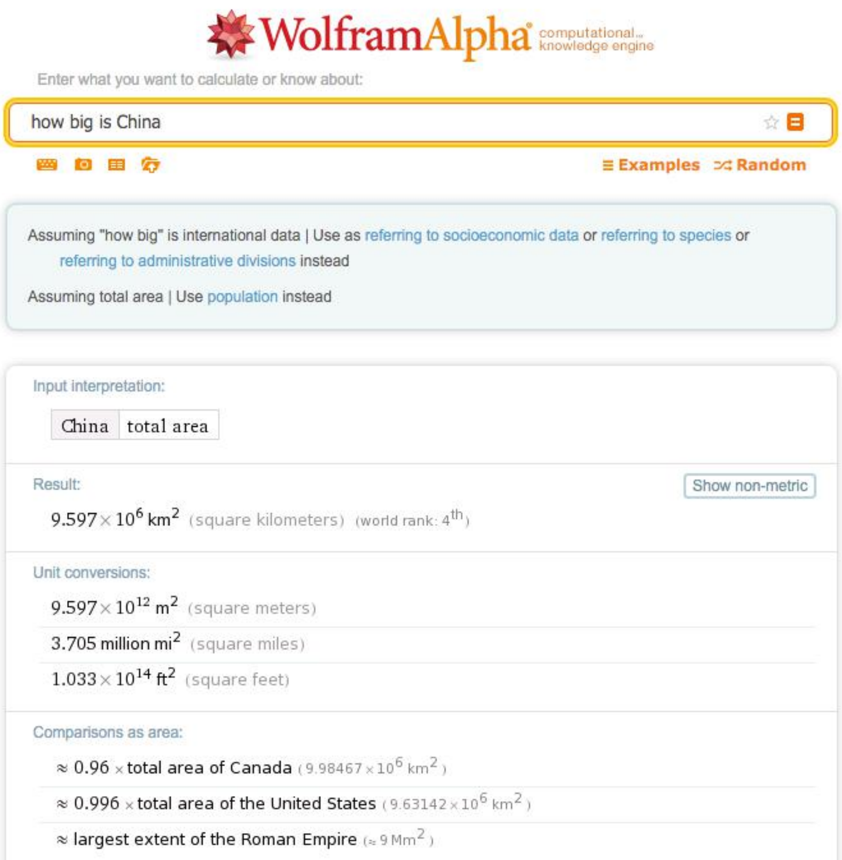
Where was Yao Ming's wife born?



KB-based QA in Search Engine



KB-based QA Applications



Basic Process

Questions
Where was Yao Ming's wife born?

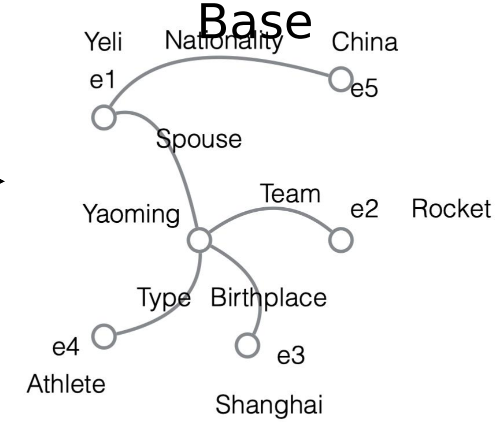
Semantic Analysis → Semantic Parsing Methods

DL-based Methods

Semantic Representation → Symbolic Representation

Distributed Representation

Matching, Reasoning ↔ Knowledge Base



Semantic Parsing

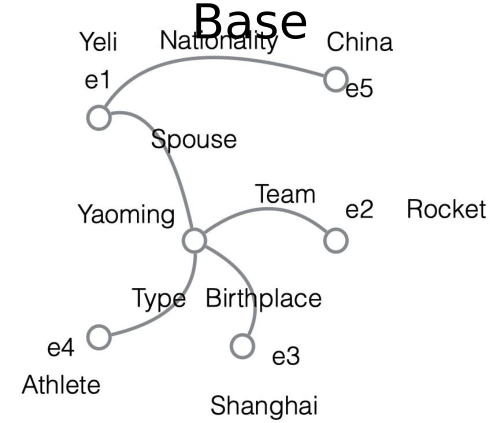
- Parse question into structured query and execute the query on knowledge graph

Where was Yao Ming's wife born?

Semantic Parsing

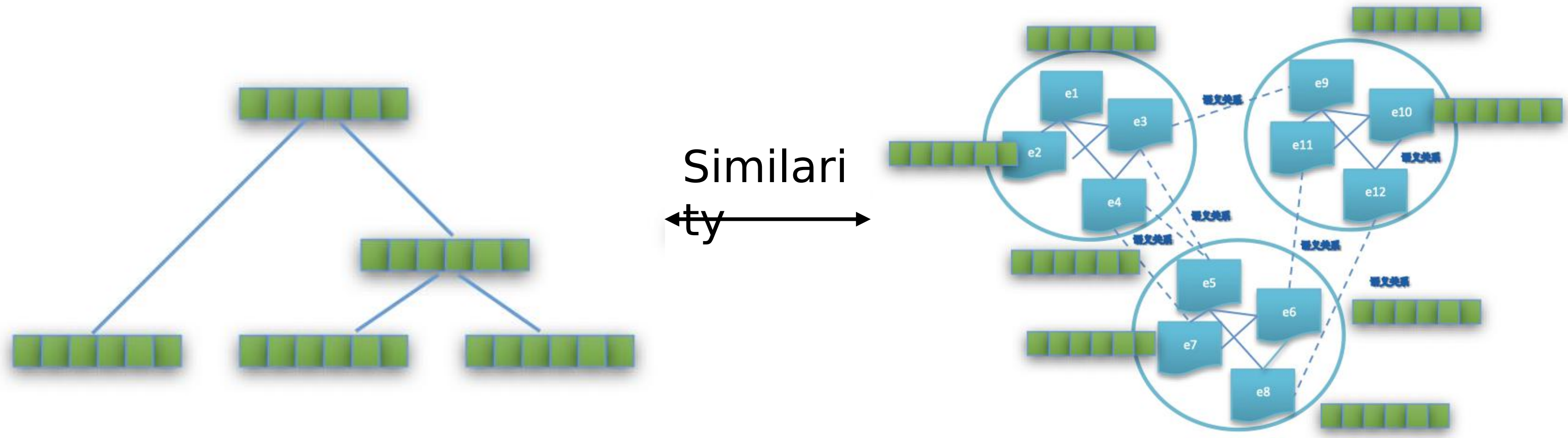
Query:
select distinct ?x
where {
    ?y Birthplace ?x.
    res: Yao Ming
Spouse ?y.
}

Knowledge Base



Deep Learning based Methods

- Encode question and alternative entities into representations and compute their similarity



Where was Yao Ming’s wife born?

Evaluation Datasets

Dataset	#train	#dev	KB	Annotation	Time
ATIS	8297	3211	ATIS	Answer	1994
Geo880	880		GeoBase	Logic form	2001
QALD-1	50	50	DBpedia	Logic form & Answer	2011
QALD-2	100	99	DBpedia & YAGO	Logic form & Answer	2012
QALD-3	100	99	DBpedia & YAGO	Logic form & Answer	2013
Free917	641	276	Freebase	Answer	2013
WebQuestion	3782	2037	Freebase	Answer	2013
WikiAnswers	2.4M	698	Reverb	Answer	2013
QALD-4	100	50	DBpedia & YAGO	Logic form & Answer	2014
QALD-5	170	59	Dbpedia &	Logic form &	2015

Outline

- Introduction to QA
- Reading Comprehension
- Open-domain QA
- KBQA
  - Definition and Dataset
  - Semantic Parsing for KBQA
  - Deep Learning for KBQA

Semantic Parsing: Constructing Structured Query

Where was Yao Ming's wife born?

λx. Wife(Yao Ming, y) ∧ Birthplace(y, x)

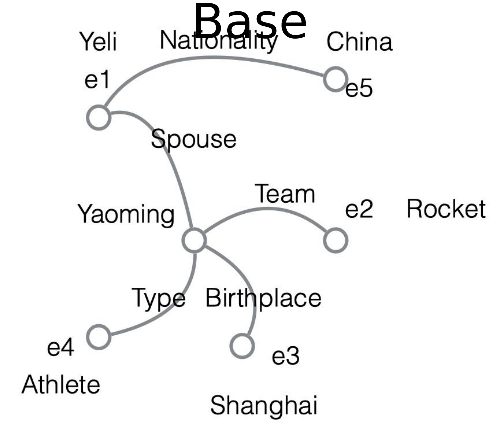
Logic Form:

- Lambda Calculus
- DCS-Tree
- …

Query:
select distinct ?x
where {
    ?y Birthplace ?x.
    res: Yao Ming
Spouse ?y.
}

SQL
SPARQL
Prolog
FunQL
…

Knowledge Base



Lambda Calculus

- Definition
  - Constants
    - Entities, Numbers, Functions
  - Logic Connectors
    - ∧, ∨, ¬, →
  - Quantification
    - ∃, ∀
  - Additional Quantifiers
    - Count, Argmax, Argmin, …
  - Lambda Expression
    - λx. state(x) ∧ borders(x, texas)

variabl
e

functio
n

entit
y

quantifi
er

a) What states border Texas

$$\lambda x.state(x) \wedge borders(x, texas)$$

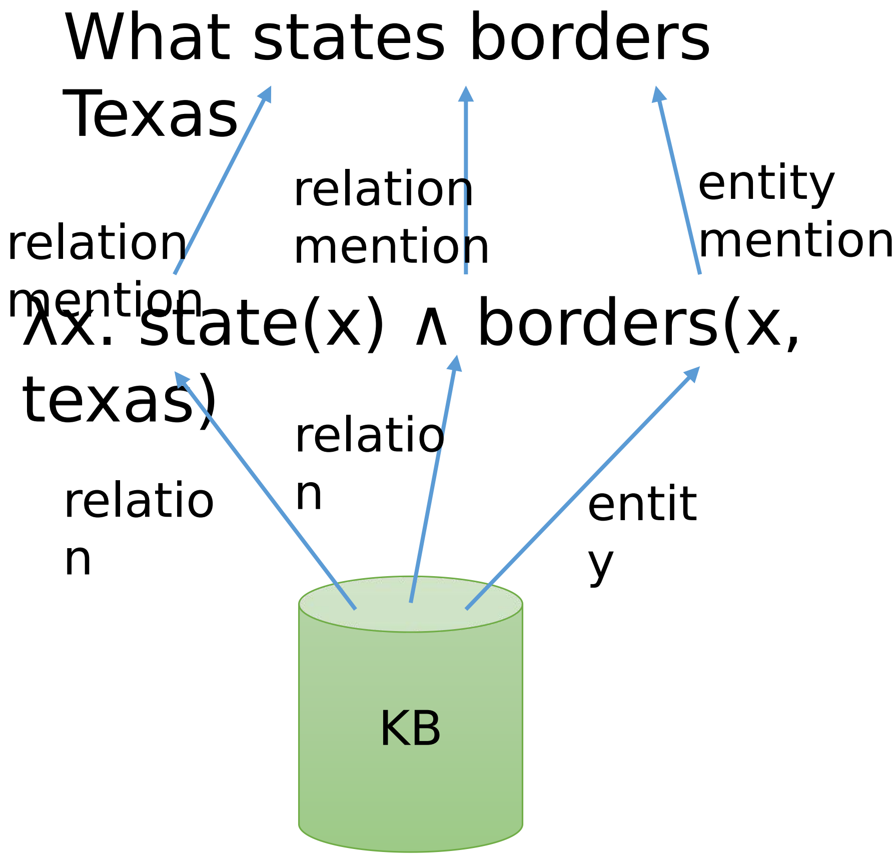
b) What is the largest state

$$\arg\max(\lambda x.state(x), \lambda x.size(x))$$

c) What states border the state that borders the most states

$$\lambda x.state(x) \wedge borders(x, \arg\max(\lambda y.state(y), \lambda y.count(\lambda z.state(z) \wedge borders(y, z))))$$

Semantic Parsing



Basic Process

Which software has been developed by organizations founded in California, USA?

↓ Key Phrase Recognition

software | developed by | organization | founded in | California

↓ Resources Linking

dbo:Software | dbr:developer | dbo:company | dbr:foundationPlace | dbo:California

↓ Semantic Composition

<dbo:Software, dbr:developer, dbo:Company> | <dbo:Company, dbr:foundationPlace, dbo:California>

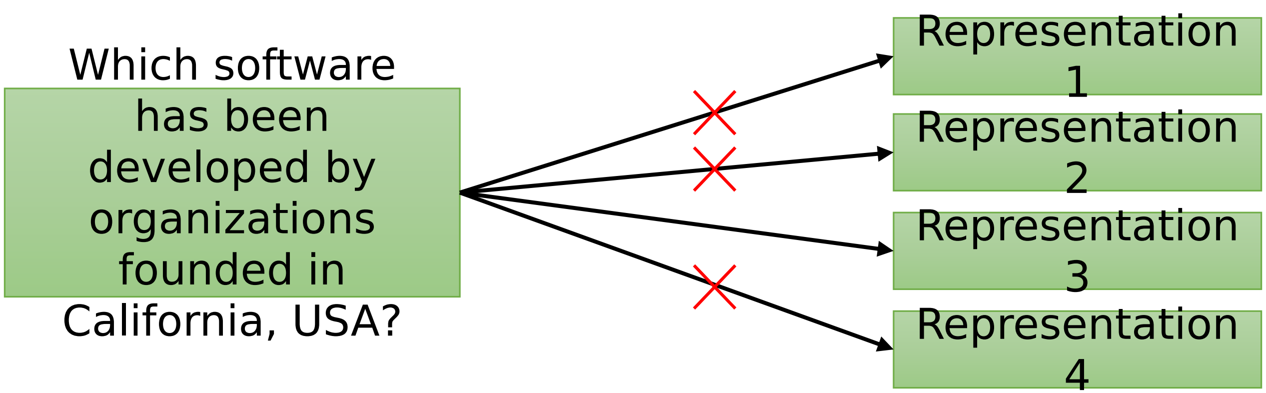
↓ Logic Form

Two Key Problems

- Resources Mapping

	Mention	KB
Class	organizati	dbo:Company
Relati on	founded in	dbr:foundationPl
Entity	California	dbo:California

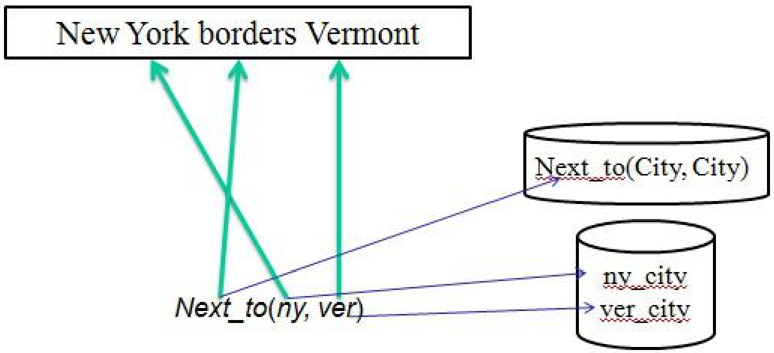
- Disambiguation



Resources Mapping: Combinatory Categorical Grammars

- Lexicon (three parts)
  - Phrase: New York
  - Category: NP
  - Resource in KB: ny

$$\text{New York} \vdash NP : ny$$
$$\text{borders} \vdash S\backslash NP/NP : \lambda x \lambda y.next\_to(y, x)$$
$$\text{Vermont} \vdash NP : vt$$



Resources Mapping: Combinatory Categorical Grammars

- Compositional Grammar

$$
\begin{array}{llcll}
X/Y : f & Y : g & \Rightarrow & X : f(g) & (>) \\
Y : g & X\backslash Y : f & \Rightarrow & X : f(g) & (<)
\end{array}
$$

$$
\begin{array}{ll}
X/Y : f \quad Y/Z : g \Rightarrow X/Z : \lambda x.f(g(x)) & (> \mathbf{B}) \\
Y\backslash Z : g \quad X\backslash Y : f \Rightarrow X\backslash Z : \lambda x.f(g(x)) & (< \mathbf{B})
\end{array}
$$

$$
\begin{array}{ccc}
\text{New York} & \text{borders} & \text{Vermont} \\
\hline
NP & (S\backslash NP)/NP & NP \\
ny & \lambda x \lambda y.next\_to(y,x) & vt \\
 & \multicolumn{2}{c}{\overline{\quad (S\backslash NP) \quad}^{>}} \\
 & \multicolumn{2}{c}{\lambda y.next\_to(y,vt)} \\
\multicolumn{3}{c}{\overline{\quad S \quad}^{<}} \\
\multicolumn{3}{c}{next\_to(ny,vt)}
\end{array}
$$

$$
\begin{array}{cccc}
\text{what} & \text{states} & \text{border} & \text{texas} \\
\hline
S/(S|NP) & S|NP/(S|NP) & S\backslash NP/NP & NP \\
\lambda f \lambda x.f(x) & \lambda f \lambda x.state(x) \wedge f(x) & \lambda y \lambda x.next\_to(x,y) & tex \\
 & \multicolumn{2}{c}{\overline{\quad S|NP/NP \quad}^{>\mathbf{B}}} & \\
 & \multicolumn{2}{c}{\lambda y \lambda x.state(x) \wedge next\_to(x,y)} & \\
 & \multicolumn{3}{c}{\overline{\quad S|NP \quad}^{>}} \\
 & \multicolumn{3}{c}{\lambda x.state(x) \wedge next\_to(x,tex)} \\
\multicolumn{4}{c}{\overline{\quad S \quad}^{>}} \\
\multicolumn{4}{c}{\lambda x.state(x) \wedge next\_to(x,tex)}
\end{array}
$$

Resources Mapping: Combinatory Categorical Grammars

- How can we map utterance to CCG Lexicon?
  - Human labeling
  - Learning from “question-logic form” pairs

1) What states border Texas?

$\lambda x.state(x) \wedge borders(x, texas)$

2) What is the largest state?

$\arg\max(\lambda x.state(x), \lambda x.size(x))$

3) What states border the state that borders the most states?

$\lambda x.state(x) \wedge borders(x, \arg\max(\lambda y.state(y), \lambda y.count(\lambda z.state(z) \wedge borders(y, z))))$

4) Utah borders Idaho.

$borders(utah, idaho)$

Utah	:=	$NP : utah$
Idaho	:=	$NP : idaho$
borders	:=	$(S\backslash NP)/NP : \lambda x.\lambda y.borders(y, x)$
states	:=	$N : \lambda x.state(x)$
major	:=	$N/N : \lambda f.\lambda x.major(x) \wedge f(x)$
population	:=	$N : \lambda x.population(x)$
cities	:=	$N : \lambda x.city(x)$
rivers	:=	$N : \lambda x.river(x)$
run through	:=	$(S\backslash NP)/NP : \lambda x.\lambda y.traverse(y, x)$
the largest	:=	$NP/N : \lambda f.\arg\max(f, \lambda x.size(x))$
river	:=	$N : \lambda x.river(x)$
the highest	:=	$NP/N : \lambda f.\arg\max(f, \lambda x.elev(x))$
the longest	:=	$NP/N : \lambda f.\arg\max(f, \lambda x.len(x))$

Question-logic form pairs

Lexico
n

Resources Mapping: Combinatory Categorical Grammars

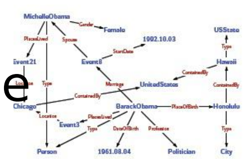
- How can we map utterance to CCG Lexicon?
  - Entity Mention (hyperlinks and redirection in Wikipedia)
  - Pattern Relation: Open IE

OpenIE

(*Barack Obama*, *was born in*, *Honolulu*)
(*Albert Einstein*, *was born in*, *Ulm*)
(*Barack Obama*, *lived in*, *Chicago*)
... 15M triples ...

was born in

Knowledge Base



(BarackObama, PlaceOfBirth, Honolulu)
(Albert Einstein, PlaceOfBirth, Ulm)
(BarackObama, PlacesLived.Location, Chicago)
... 600M triples ...

PlaceOfBirth

Disambiguation: Probabilistic Combinatory Categorical Grammars

- Natural Language is expressed in a more complex way
- Example: Which software has been developed by organizations founded in California, USA?
  - Key phrase recognition: { California }, { California, USA }
  - Resources Linking:
    - California: {California_State}, {California_Film}
    - founded: {foundationPlace},{founder}
    - developed by: {developer}
  - Semantic Composition:
    - {dbo:Software, dbr:developer, dbo:Company}
    - {dbo:Software, dbr:foundationPlace, dbo:Company}

Disambiguation: Probabilistic Combinatory Categorical Grammars

- Probabilistic Combinatory Categorical Grammars (PCCG)

$$L = \arg\max P(L|S;\theta) = \arg\max \sum_T P(L,T|S;\theta)$$

$$P(L,T|S;\theta) = \frac{\exp f(L,T,S|\theta)}{\sum_{(L,T)} \exp f(L,T,S|\theta)}$$

- L: predicted logic form
- T: CCG parse
- S: sentence
- θ: features

Outline

- Introduction to QA
- Reading Comprehension
- Open-domain QA
- KBQA
  - Definition and Dataset
  - Semantic Parsing for KBQA
  - Deep Learning for KBQA

Employing DL to improve KBQA

- DL-enhanced Methods

Key Phrase Detection

↓

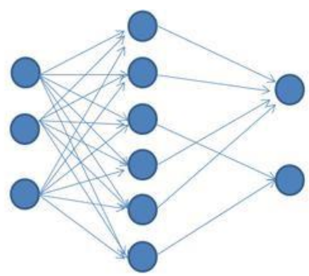
Resource Linking

↓

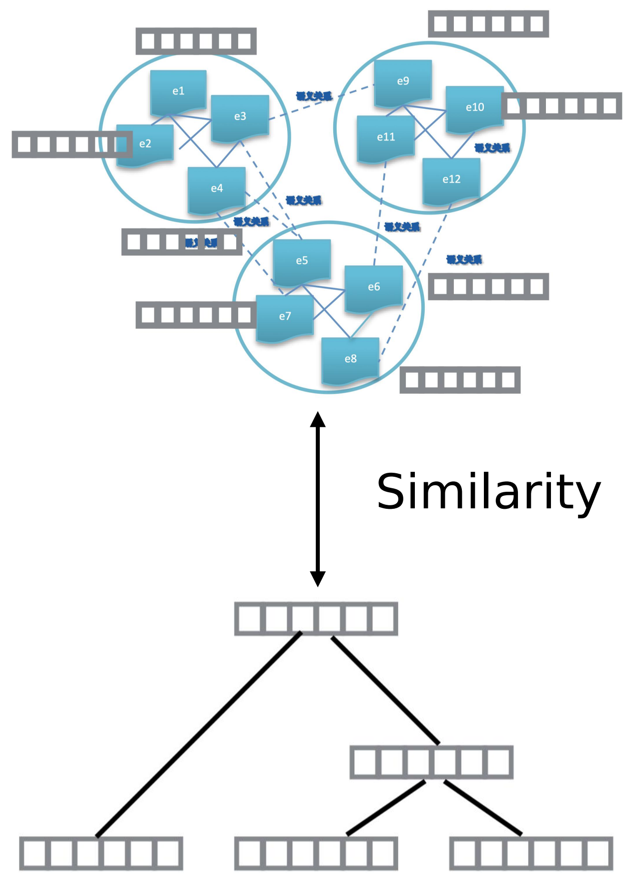
Semantic Composition

↓

Logic form



- End2End Models

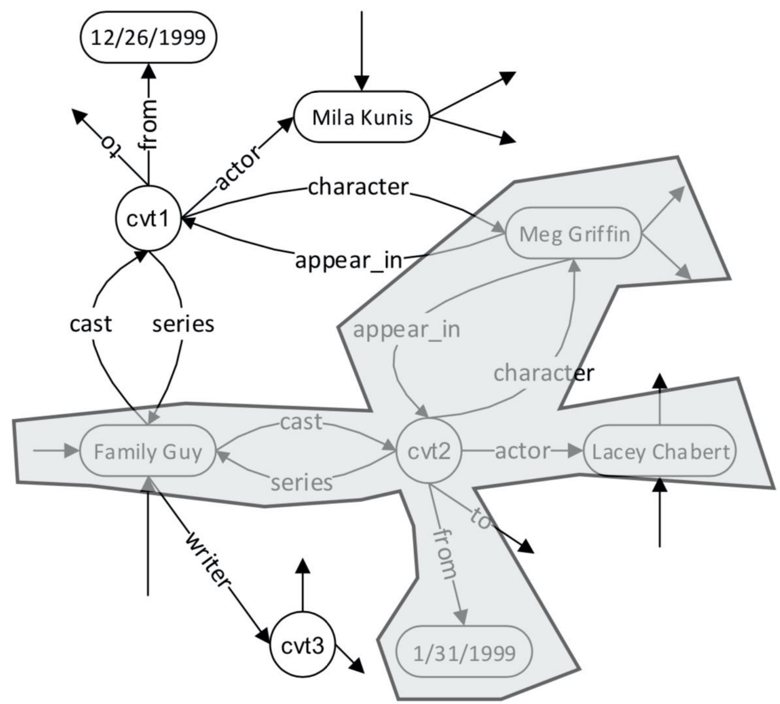


DL-enhanced Methods: Staged Query Graph Generation

- Generating structured queries that can be executed on knowledge bases

Who first voiced
Meg on Family Guy?

$$\mathrm{argmin}(\lambda x.\mathrm{Actor}(x, \mathrm{Family\_Guy}) \wedge \mathrm{Voice}(x, \mathrm{Meg\_Griffin}), \lambda x.\mathrm{casttime}(x))$$

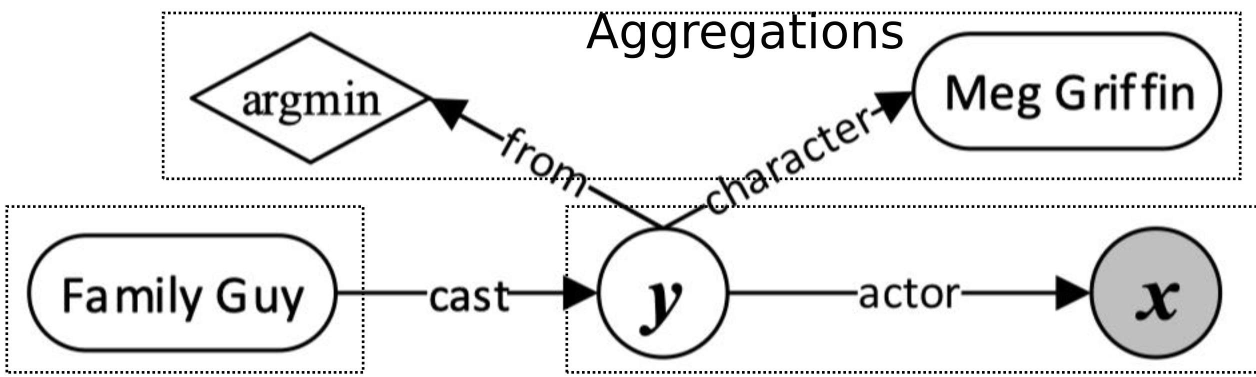


Yih et al. Semantic Parsing via Staged Query Graph Generation: Question Answering with Knowledge Base. In Proceedings of ACL 2015

DL-enhanced Methods: Staged Query Graph Generation

- Formalize query graph generation as a search problem
- Generate all possible query graphs and score them
- Example: Who first voiced Meg on Family Guy?

3. Constraints & Aggregations



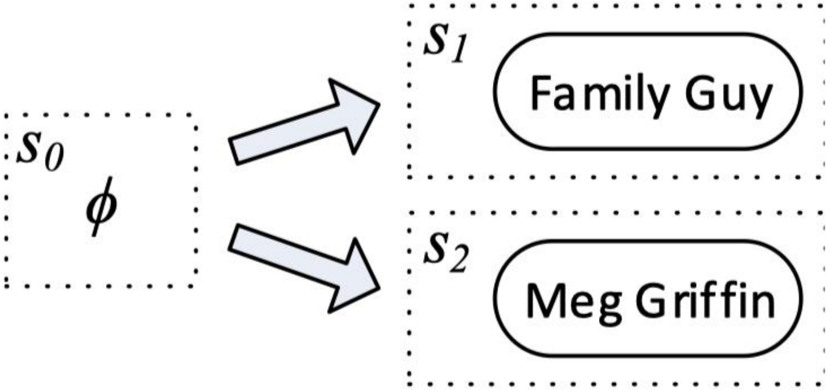
1. Topic entity

2. Core Inferential Chain

Yih et al. Semantic Parsing via Staged Query Graph Generation: Question Answering with Knowledge Base. In Proceedings of ACL 2015

Linking Topic Entity

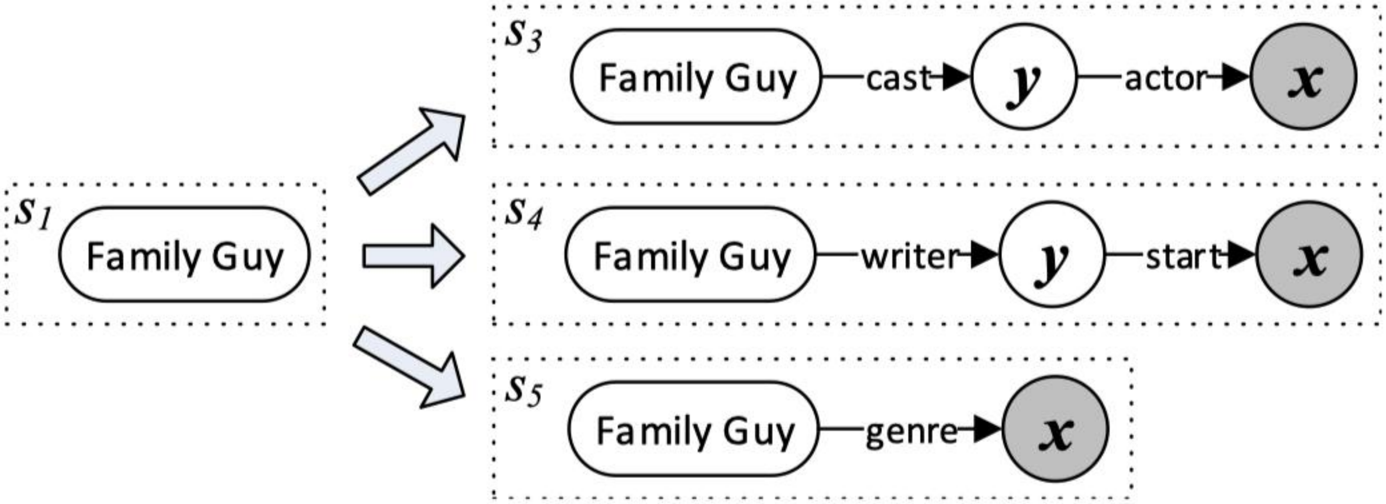
- Step 1: Lexicon-based candidate generation (anchor texts, redirect pages and others)
- Step 2: Selecting entities by using frequency counts (Top 10)
- Example: Who first voiced Meg on Family Guy?



Yih et al. Semantic Parsing via Staged Query Graph Generation: Question Answering with Knowledge Base. In Proceedings of ACL 2015

Identify Core Inferential Chain

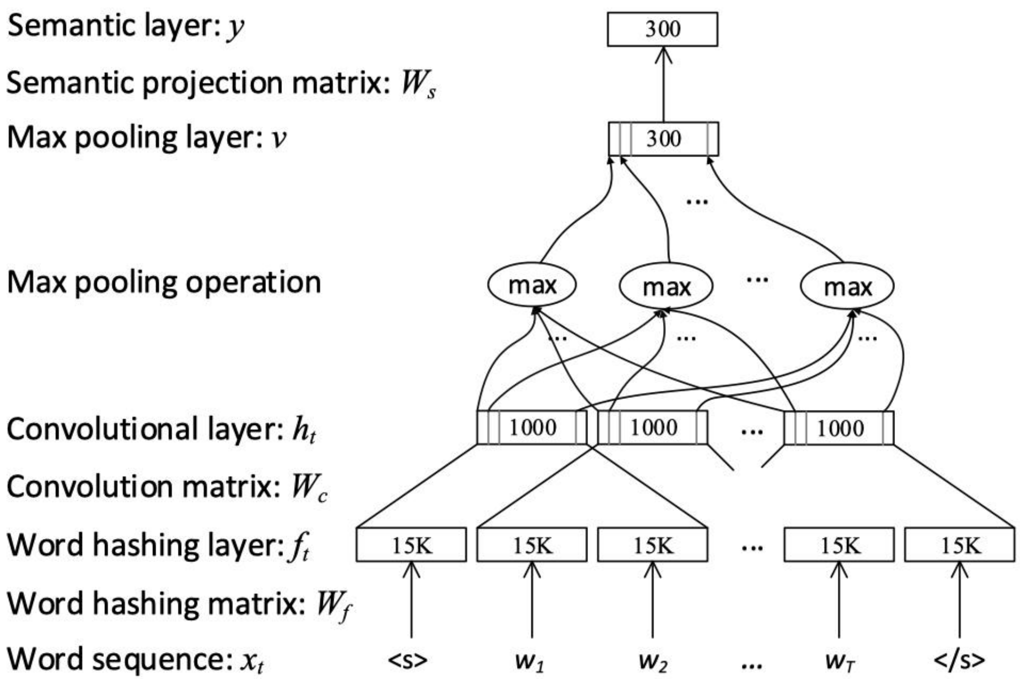
- Explore legitimate predicate sequences that can start from a top entity
  - 2 steps if y is a CVT-node
  - 1 step if y is a non-CVT-node
- Example: Who first voiced Meg on Family Guy?



Yih et al. Semantic Parsing via Staged Query Graph Generation: Question Answering with Knowledge Base. In Proceedings of ACL 2015

Question/Inferential Chain Encoder

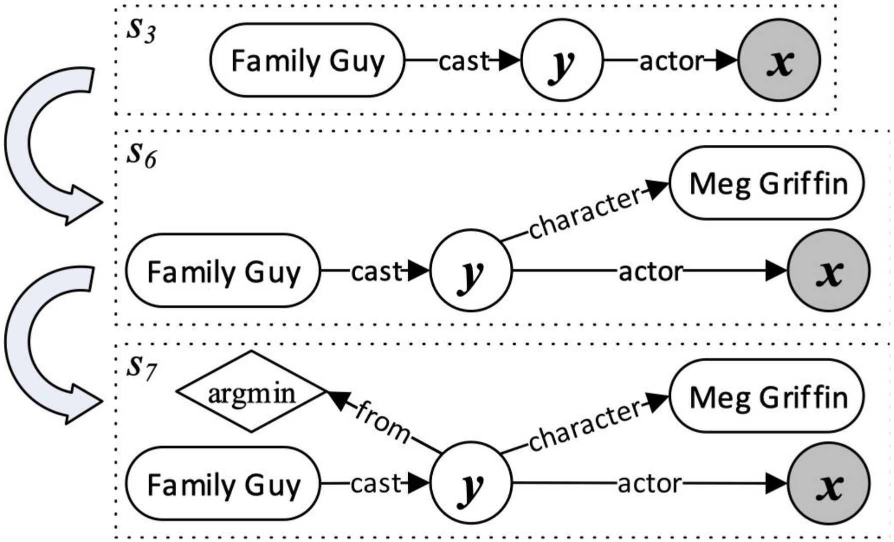
- Break word into a vector of letter-trigrams
  - who → #wh, who, ho#.
- Apply CNN, Max-Pooling, FFN on letter-trigrams



Yih et al. Semantic Parsing via Staged Query Graph Generation: Question Answering with Knowledge Base. In Proceedings of ACL 2015

Augmenting Constraints & Aggregations

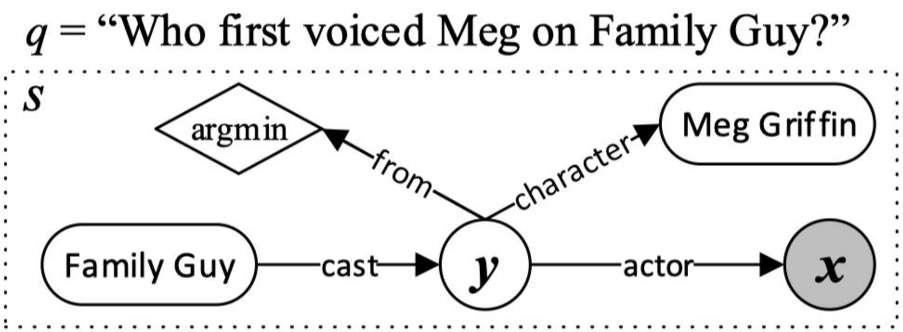
- Augmenting Constraints: add entities that are neighbor nodes of x or y, and appears in question
- Aggregations: can be added if certain keywords like “first” or “latest” occur in question
- Example: ' Family Guy?



Yih et al. Semantic Parsing via Staged Query Graph Generation: Question Answering with Knowledge Base. In Proceedings of ACL 2015

Rank Generated Graphs

- Log Linear Model
- Features
  - Topic Entity: Entity Linking Score
  - Core Inferential Chain: Relation Matching Score (NN-based model)
  - Constraints: keyword and entity matching



(1) EntityLinkingScore(FamilyGuy, "Family Guy") = 0.9
(2) PatChain("who first voiced meg on <e>", cast-actor) = 0.7
(3) QuesEP(q, "family guy cast-actor") = 0.6
(4) ClueWeb("who first voiced meg on <e>", cast-actor) = 0.2
(5) ConstraintEntityWord("Meg Griffin", q) = 0.5
(6) ConstraintEntityInQ("Meg Griffin", q) = 1
(7) AggregationKeyword(argmin, q) = 1
(8) NumNodes(s) = 5
(9) NumAns(s) = 1

Yih et al. Semantic Parsing via Staged Query Graph Generation: Question Answering with Knowledge Base. In Proceedings of ACL 2015

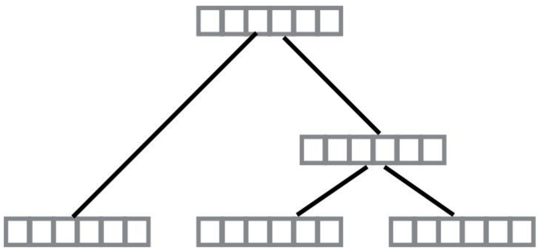
Results

- Bechmark: WebQuestions
  - 5,810 Q-A pairs from google query log
  - Download: http://nlp.stanford.edu/software/sempre/

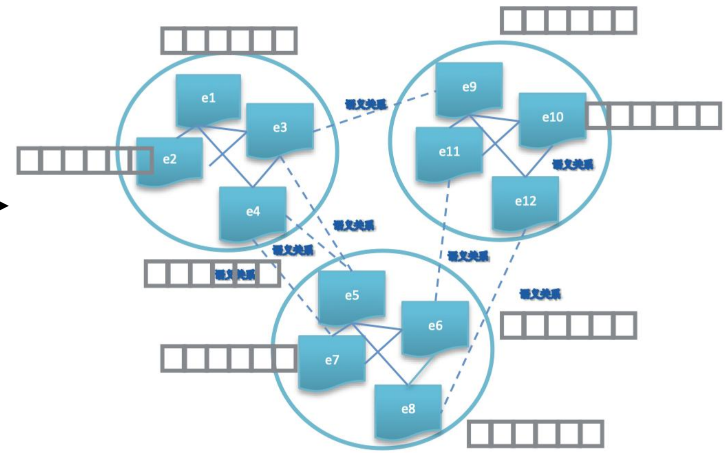
Method	Prec.	Rec.	F_1
(Berant et al., 2013)	48.0	41.3	35.7
(Bordes et al., 2014b)	-	-	29.7
(Yao and Van Durme, 2014)	-	-	33.0
(Berant and Liang, 2014)	40.5	46.6	39.9
(Bao et al., 2014)	-	-	37.5
(Bordes et al., 2014a)	-	-	39.2
(Yang et al., 2014)	-	-	41.3
(Wang et al., 2014)	-	-	45.3
Our approach – STAGG	52.8	60.7	**52.5**

Yih et al. Semantic Parsing via Staged Query Graph Generation: Question Answering with Knowledge Base. In Proceedings of ACL 2015

End2End Based KBQA



Semant
ic
Similari
ty



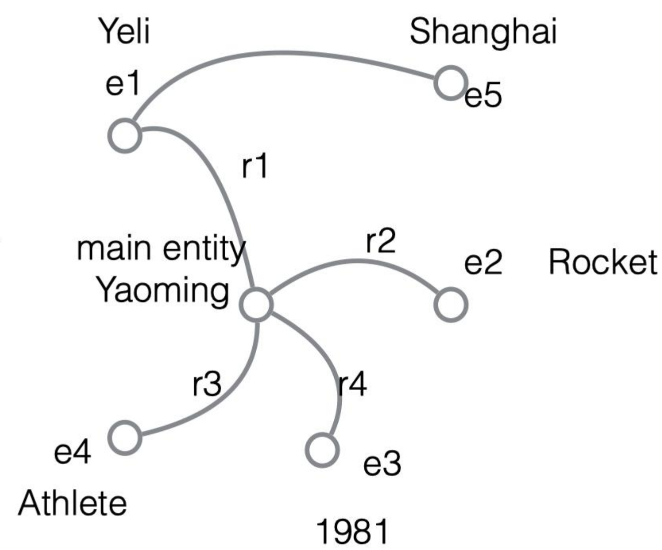
Where was Yao Ming’s wife born?

- How to represent the input question?
- How to represent the entity/relation in KB?
- How to match question and entities/relations?

Basic End2End QA System

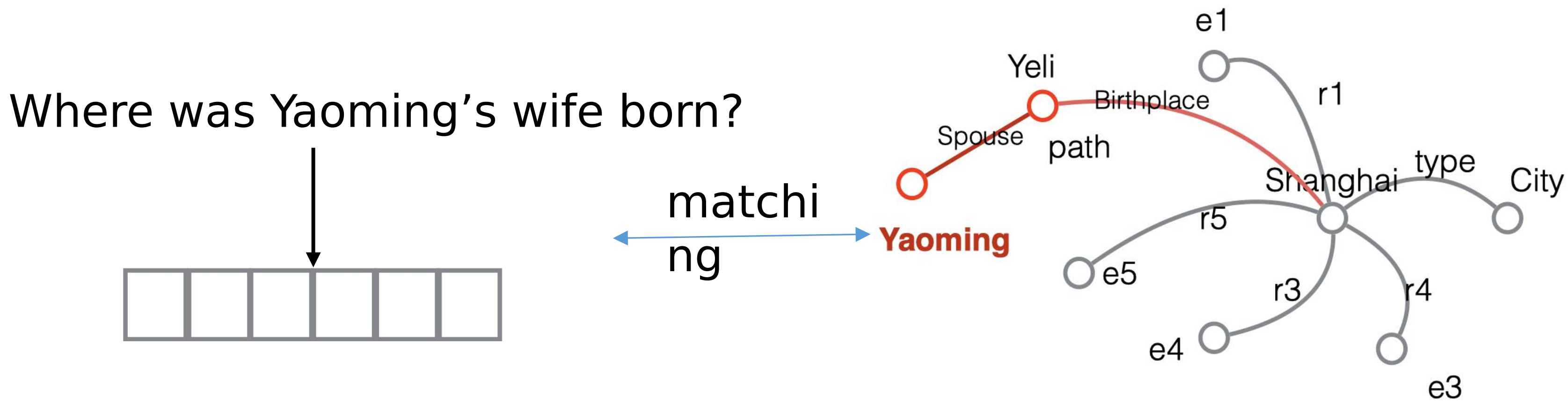
- Step1: Candidate Generation
  - Finding main entity in the question by Entity Linking
  - All entities around the main entity in KB are candidates
- Step2: Candidate Ranking

Where was **Yao Ming's** wife born?



Bordes et al. Question Answering with Subgraph Embedding, In Proceedings of EMNLP 2014

Basic End2End QA System



- Scoring function:

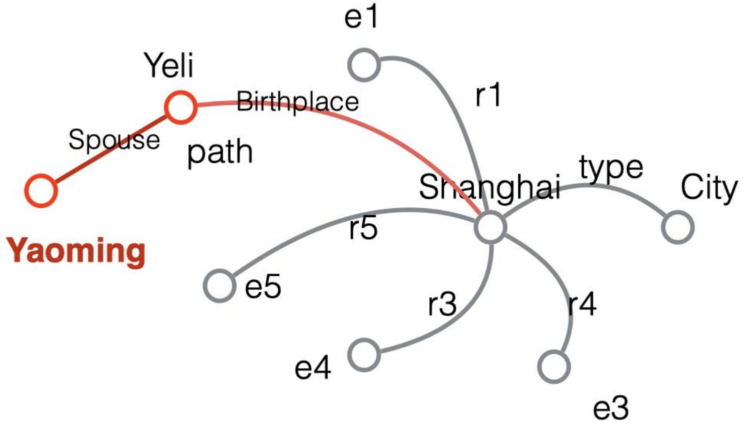
$$S(q,a) = f(q)^{\mathrm{T}} g(a)$$
$$a = \arg\max S(q,a)$$

- Training Objective: $L = \max(0,\ 1 - S(q,a_{+}) + S(q,a_{-}))$
- Question Embedding: $f(q) = \sum_{w \in q} E(w)$

Bordes et al. Question Answering with Subgraph Embedding, In Proceedings of EMNLP 2014

Basic End2End QA System

- Answer Embedding: Sum
  - Entity Representation
    - $E(\text{Shanghai})$
  - Path Representation
    - $E(\text{Yaoming}) + E(\text{Spouse})$
    $+E(\text{Birthplace}) + E(\text{Shanghai},$
  - Subgraph Representation
    - Sum of embeddings of answer's neighbors and the relations between answer and neighbors
    - $E'(\text{Yeli}) + E'(\text{Birthplace}) + E'(\text{e1}) + E'(\text{r1}) + E'(\text{city}) + E'(\text{type}) + \dots$
    - Different embedding for entities/relations in path and subgraph



Bordes et al. Question Answering with Subgraph Embedding, In Proceedings of EMNLP 2014

Basic End2End QA Systems

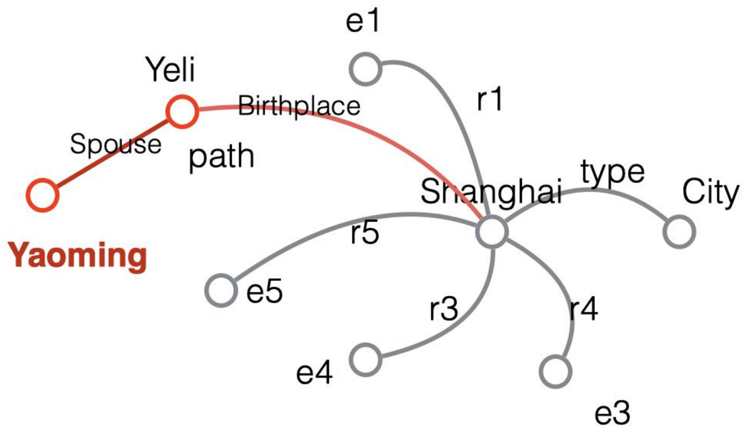
- Results on WebQuestions

Method	P@1 (%)	F1 (Berant)	F1 (Yao)
Baselines			
(Berant et al., 2013)	–	31.4	–
(Bordes et al., 2014b)	31.3	29.7	31.8
(Yao and Van Durme, 2014)	–	33.0	42.0
(Berant and Liang, 2014)	–	39.9	43.0
Our approach			
Subgraph & $\mathcal{A}(q) = C_2$	**40.4**	39.2	43.2
Ensemble with (Berant & Liang, 14)	–	**41.8**	**45.7**

Bordes et al. Question Answering with Subgraph Embedding, In Proceedings of EMNLP 2014

Multi-Column CNN

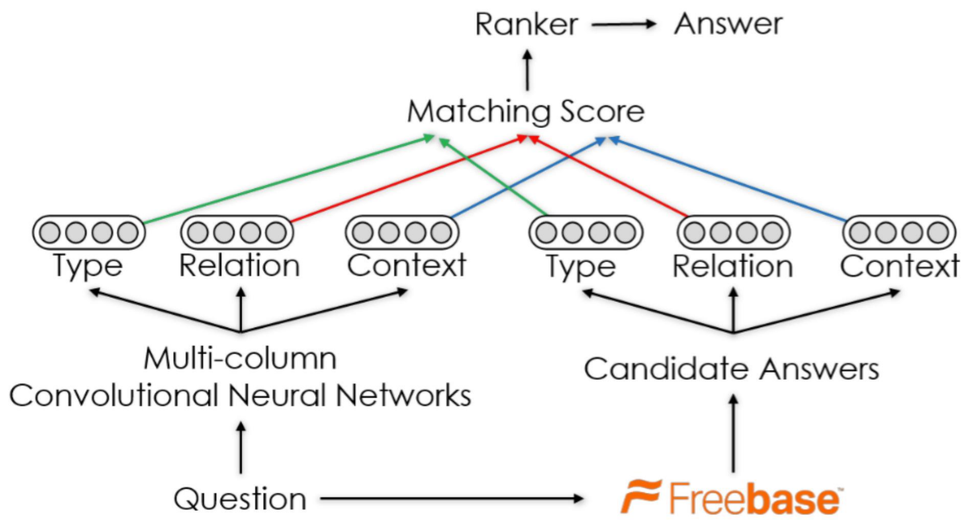
- Problem:
  - Question representation should be different according to different focused views
  - Example:
    - Where was Yaoming's **wife born**?
    - Answer Path: Spouse—>Birthplace
    - **Where** was Yaoming's wife born?
    - Answer Type: City

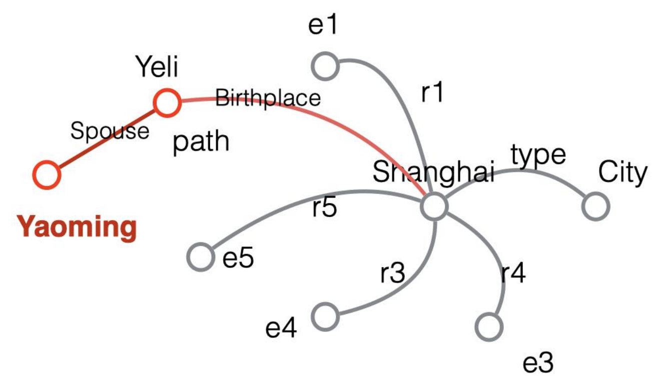


Dong et al. Question Answering over Freebase with Multi-Column Convolutional Neural Network, In Proceedings of ACL 2015

Multi-Column CNN

- Employ different CNNs to consider different views





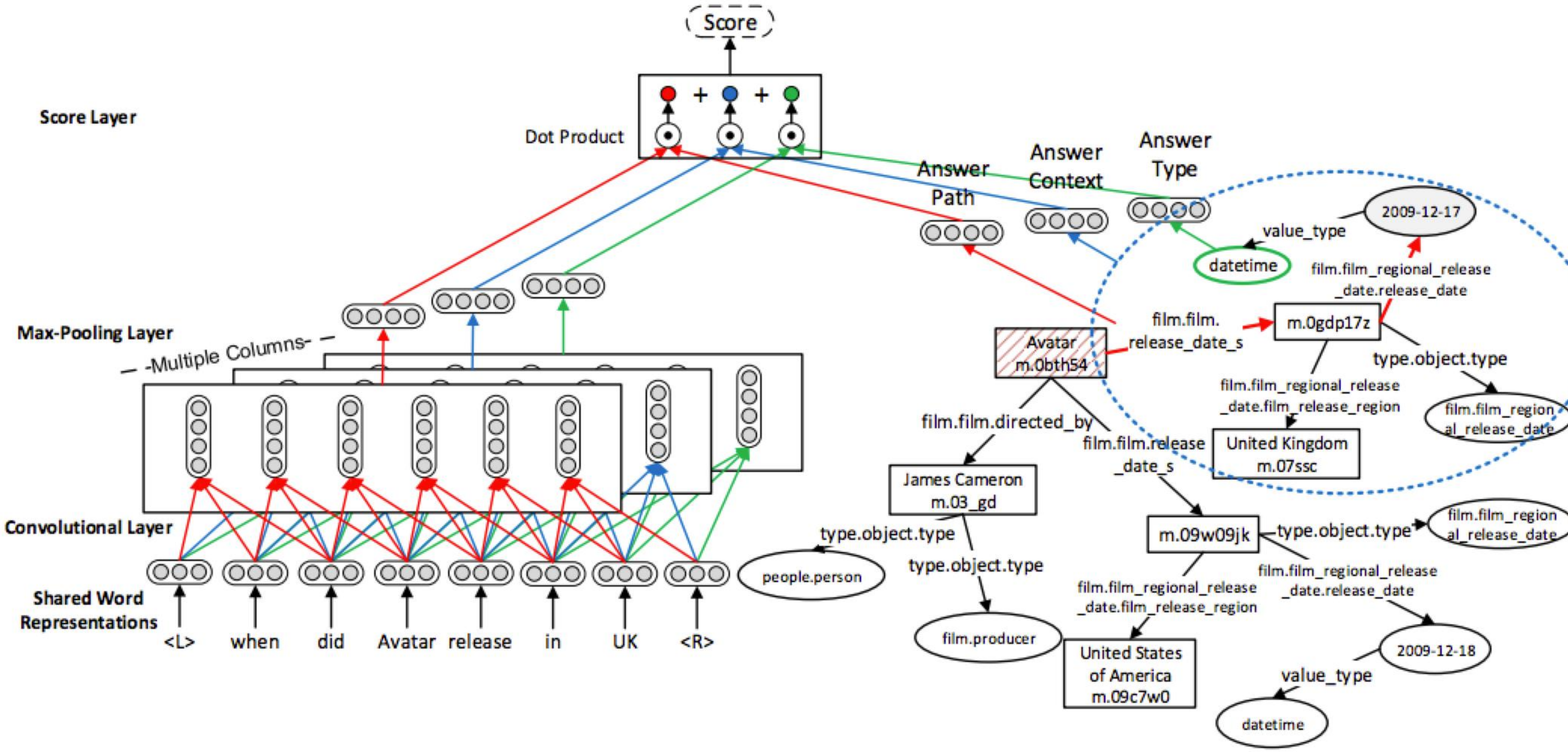
$$S(q,a) = \underbrace{\mathbf{f}_1(q)^{\mathsf{T}}\mathbf{g}_1(a)}_{answer\ path} + \underbrace{\mathbf{f}_2(q)^{\mathsf{T}}\mathbf{g}_2(a)}_{answer\ context} + \underbrace{\mathbf{f}_3(q)^{\mathsf{T}}\mathbf{g}_3(a)}_{answer\ type}$$

Answer Type: City
Answer Context: (e1, r1), (e3, r4), (e4, r3), (e5, r5), (Yeli, Birthplace)
Answer Path: Spouse, Birthplace

Dong et al. Question Answering over Freebase with Multi-Column Convolutional Neural Network, In Proceedings of ACL 2015

Multi-Column CNN

- Example: when did Avatar release in UK?



Dong et al. Question Answering over Freebase with Multi-Column Convolutional Neural Network, In Proceedings of ACL 2015

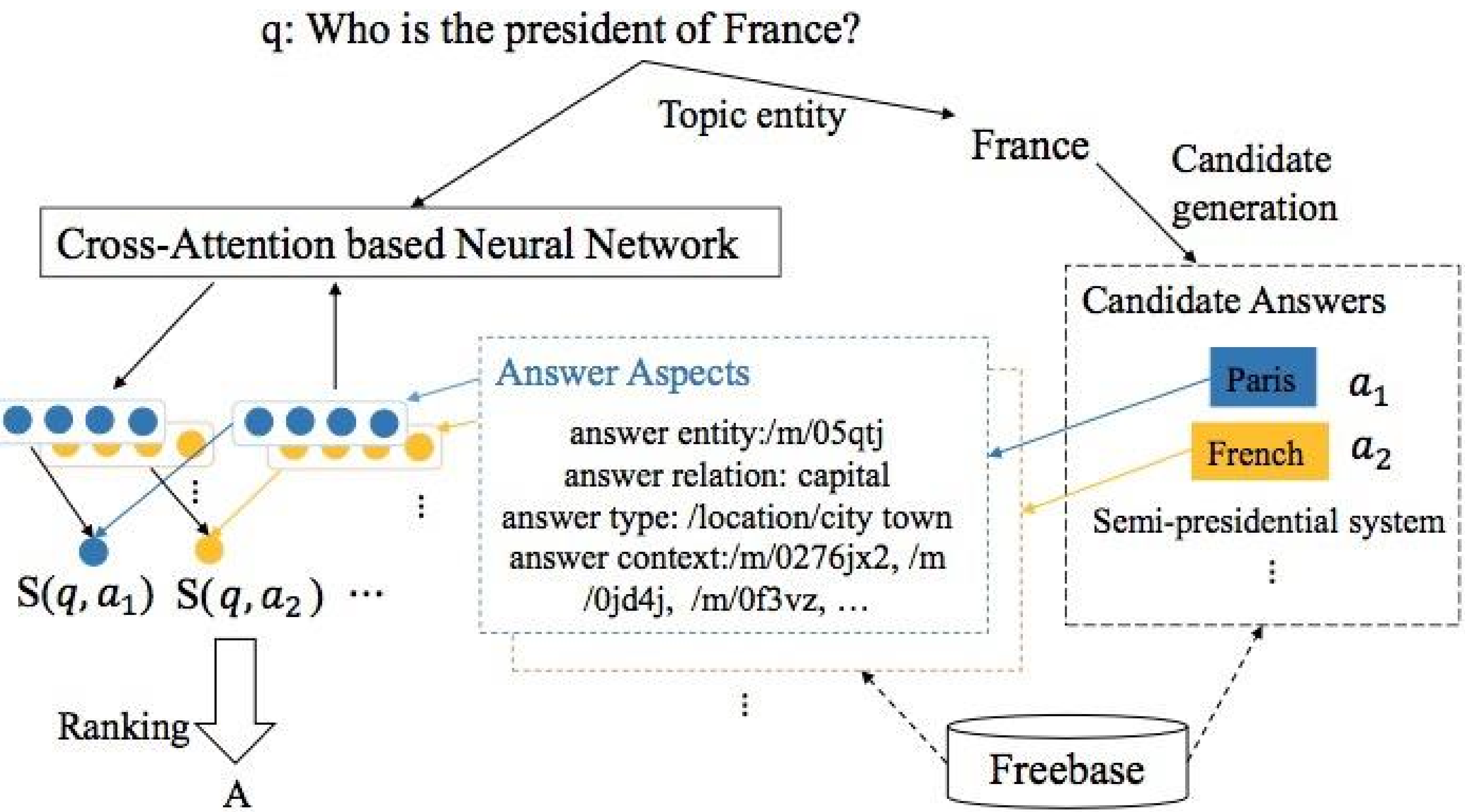
Multi-Column CNN

- Results

Method	F1	P@1
(Berant et al., 2013)	31.4	-
(Berant and Liang, 2014)	39.9	-
(Bao et al., 2014)	37.5	-
(Yao and Van Durme, 2014)	33.0	-
(Bordes et al., 2014a)	39.2	40.4
(Bordes et al., 2014b)	29.7	31.3
MCCNN (our)	**40.8**	**45.1**

Dong et al. Question Answering over Freebase with Multi-Column Convolutional Neural Network, In Proceedings of ACL 2015

Employing RNN with Attention



Hao et al. An End-to-End Model for Question Answering over Knowledge Base with Cross-Attention Combining Global Knowledge, In Proceedings of ACL 2017

Cross Attention

A-Q Attention

$$S(q, e_i) = h(e_i, \sum_j \alpha_{ij} h_j)$$

$$\alpha_{ij} = \frac{\exp a_{ij}}{\sum_j \exp a_{ij}}$$

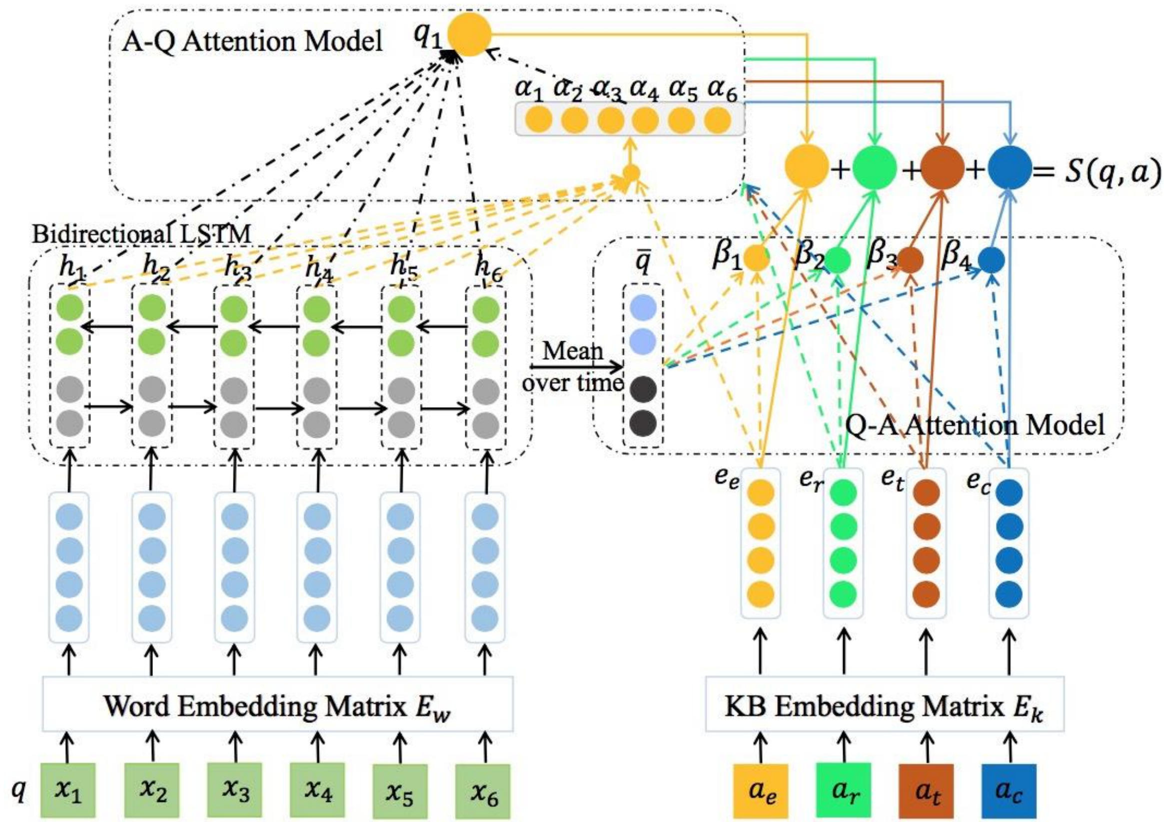
$$a_{ij} = f(W[h_j; e_i] + b)$$

Q-A Attention

$$S(q, a) = \sum_{e_i \in \{e_e, e_r, e_t, e_c\}} \beta_{e_i} S(q, e_i)$$

$$\beta_{e_i} = \frac{\exp w_{e_i}}{\sum_{e_j} \exp w_{e_j}}$$

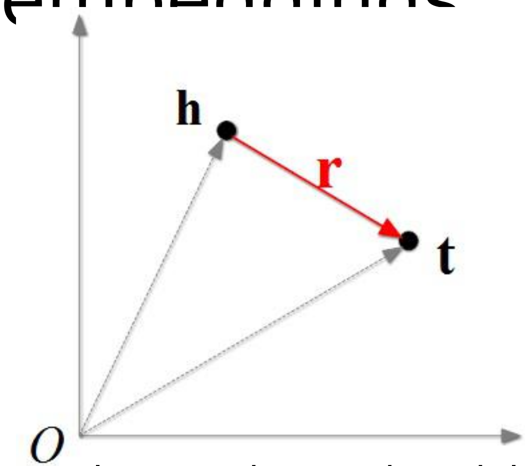
$$w_{e_j} = f(W'[e_i, \frac{1}{n} \sum_k h_k])$$



Hao et al. An End-to-End Model for Question Answering over Knowledge Base with Cross-Attention Combining Global Knowledge, In Proceedings of ACL 2017

Combining Global Knowledge Information

- The knowledge representations in previous methods are limited in training data
  - Only the embeddings of knowledge in training set are trained
  - OOV problems
- Incorporating pre-trained knowledge embeddings



$$\mathbf{h} + \mathbf{r} = \mathbf{t}$$

$$f(h, r, t) = |h + r - t|_2$$

$$\sum_{(h,r,t)\in\Delta} \sum_{(h',r,t')\notin\Delta} [\gamma + f(h, r, t) - f(h', r, t')]_+$$

Hao et al. An End-to-End Model for Question Answering over Knowledge Base with Cross-Attention Combining Global Knowledge, In Proceedings of ACL 2017

Multi-task Learning

- Question Answering

$$L_{q,a,a'} = [1 - S(q, a_+) + S(q, a_-)]_+$$

- Knowledge Embedding

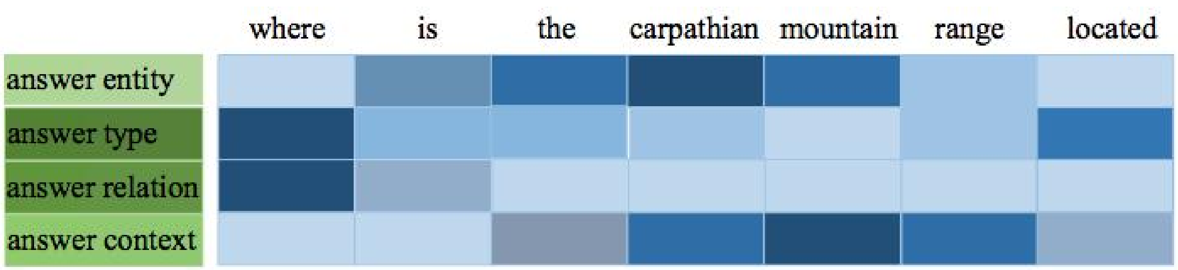
$$L_k = \sum_{(h,r,t)\in\Delta} \sum_{(h',r,t')\notin\Delta} [\gamma + f(h, r, t) - f(h', r, t')]_+$$

Hao et al. An End-to-End Model for Question Answering over Knowledge Base with Cross-Attention Combining Global Knowledge, In Proceedings of ACL 2017

Results

Methods	Avg F_1
Bordes et al., 2014b	29.7
Bordes et al., 2014a	39.2
Yang et al., 2014	41.3
Dong et al., 2015	40.8
Bordes et al., 2015	42.2
our approach	**42.9**

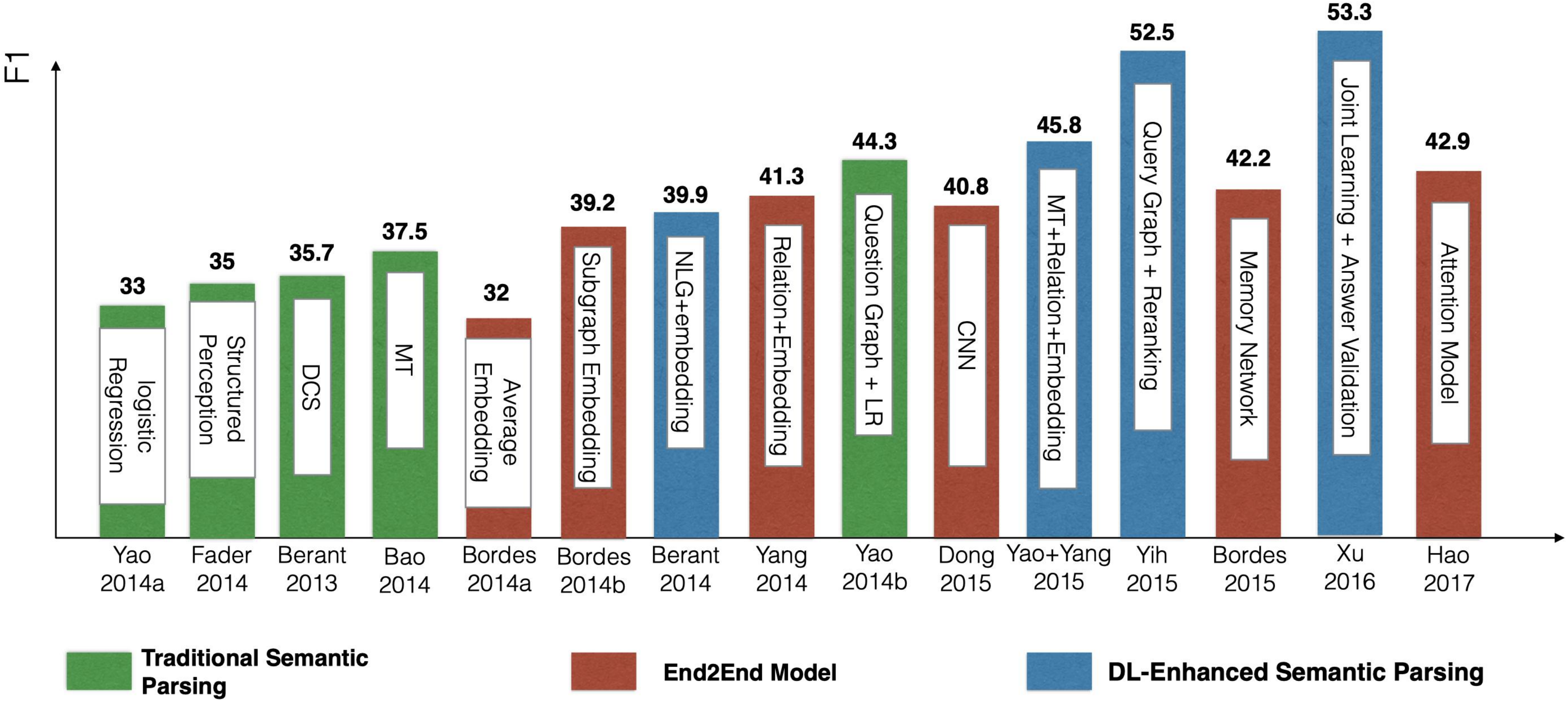
Methods	Avg F_1
LSTM	38.2
Bi_LSTM	39.1
Bi_LSTM+A-Q-ATT	41.6
Bi_LSTM+C-ATT	41.8
Bi_LSTM+GKI	40.4
Bi_LSTM+A-Q-ATT+GKI	42.6
Bi_LSTM+C-ATT+GKI	42.9



entity: Slovakia
type: /location/country
relation: partially/containedby
context: /m/04dq9kf, /m/01mp…

Hao et al. An End-to-End Model for Question Answering over Knowledge Base with Cross-Attention Combining Global Knowledge, In Proceedings of ACL 2017

Model Comparison on WebQuestion



Summary

- Reading Comprehension
  - Cloze test, Extractive RC, Multiple-choice
  - Bidirectional attention
  - Multi-paragraph, Shared-norm
- Open-domain QA
  - Retriever + Reader
- KBQA
  - Semantic Parsing, CCG
  - DL-enhanced KBQA
  - End2End KBQA

Reading Material

a. Reading Comprehension

SQuAD: 100,000+ Questions for Machine Comprehension of Text. EMNLP 2016 [link]

Bidirectional Attention Flow for Machine Comprehension. ICLR 2017 [link]

Simple and Effective Multi-Paragraph Reading Comprehension. ACL 2018 [link]

b. Open-domain QA

Reading Wikipedia to Answer Open-Domain Questions. ACL 2017 [link]

Open Domain Question Answering Using Early Fusion of Knowledge Bases and Text. EMNLP 2018 [link]

c. KBQA

Question Answering with Subgraph Embedding. EMNLP 2014 [link]

Semantic Parsing via Staged Query Graph Generation: Question Answering with Knowledge Base. ACL 2015 [link]

For reading material recommendation of this course, please refer to our github

Reading Material

d. Other Topics

(Multi-hop) Self-Assembling Modular Networks for Interpretable Multi-Hop Reasoning. EMNLP 2019 [link]

(Symbolic) Neural symbolic Reader: scalable integration of distributed and symbolic representations for reading comprehension. ICLR 2020 [link]

(Adversial) Adversarial Examples for Evaluating Reading Comprehension Systems. EMNLP 2017 [link]

(PIQA) Phrase-indexed question answering: A new challenge for scalable document comprehension. EMNLP 2018 [link]

(Common Sense) Graph-Based Reasoning over Heterogeneous External Knowledge for Commonsense Question Answering. [link]

(CQA) SDNet: Contextualized Attention-based Deep Network for Conversational Question Answering. [link]

For reading material recommendation of this course, please refer to our github.

Q&A

THUNLP