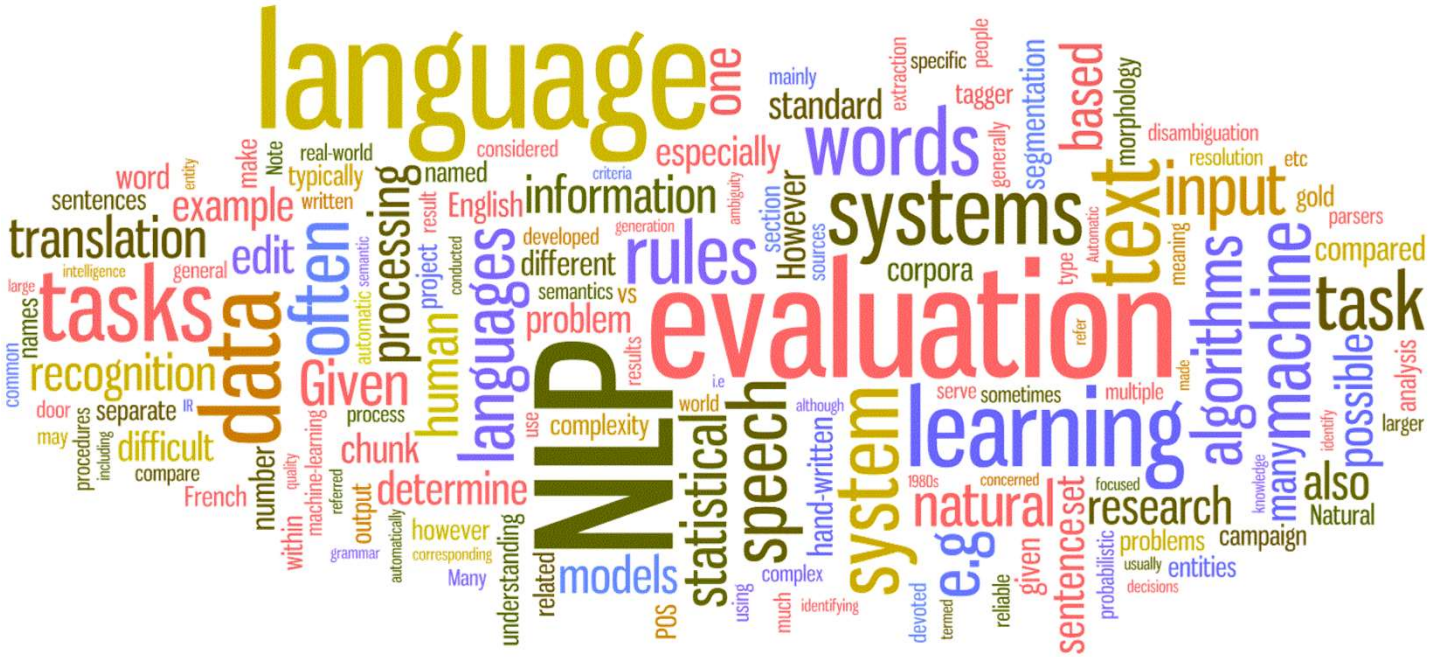


Lecture 15: Question Answering

Xu Ruifeng

Harbin Institute of Technology, Shenzhen

Overview

Today we'll be learning about **Question Answering (QA) systems.**

Plan for today:

- What is question answering?
- IR-based question answering
- Knowledge-based question answering
- Hybrid approaches (IBM Watson)
- QA Task: Machine Reading Comprehension

Question Answering

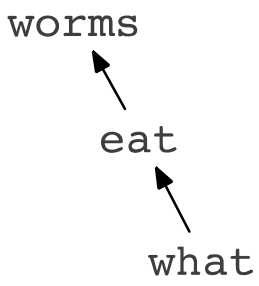
What is Question Answering?

Question Answering

- One of the oldest NLP tasks (punched card systems in 1961)

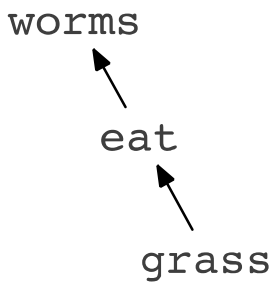
Question:

What do worms eat?

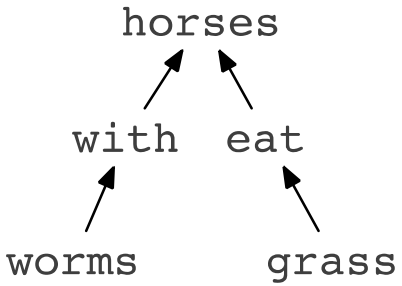


Potential Answers:

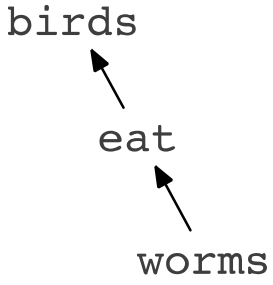
Worms eat grass



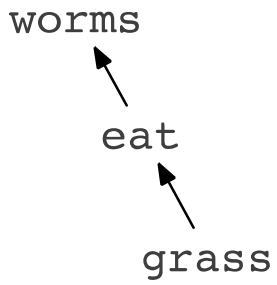
Horses with worms eat grass



Birds eat worms



Grass is eaten by worms



Simmons, Klein, McConlogue. 1964. Indexing and Dependency Logic for Answering English Questions. American Documentation 15:30, 196-204

Question Answering: IBM's Watson

- Won Jeopardy on February 16, 2011!

WILLIAM WILKINSON'S
"AN ACCOUNT OF THE PRINCIPALITIES OF WALLACHIA AND MOLDOVIA"
INSPIRED THIS AUTHOR'S
MOST FAMOUS NOVEL

Bram Stoker

Question Answering: Apple's Siri

Question Answering: Baidu

Types of Questions in Modern Systems

- Factoid questions
  - *中国的首都是哪个城市?*
  - *深圳在哪?*
  - *2019年中国的GDP是多少?*
  - *目前中国的人口有多少?*
- Complex (narrative) questions:
  - *基于mRNA研发的疫苗效果怎么样?*
  - *国外学者们如何看待中国抗疫的成果?*

Commercial systems: factoid questions

周杰伦的妻子是谁	昆凌
阿里巴巴市值多少	7034.66亿美元
特斯拉创始人的出生地是哪里?	南非
深圳在哪里	广东
中国的首都是哪个城市	北京
中国GDP预计哪一年超过美国	2030年

Paradigms for QA

- IR-based approaches
  - IBM Watson
  - Google
  - Baidu
  - Bing
- Knowledge-based and Hybrid approaches
  - IBM Watson
  - Apple Siri
  - Wolfram Alpha
  - True Knowledge Evi

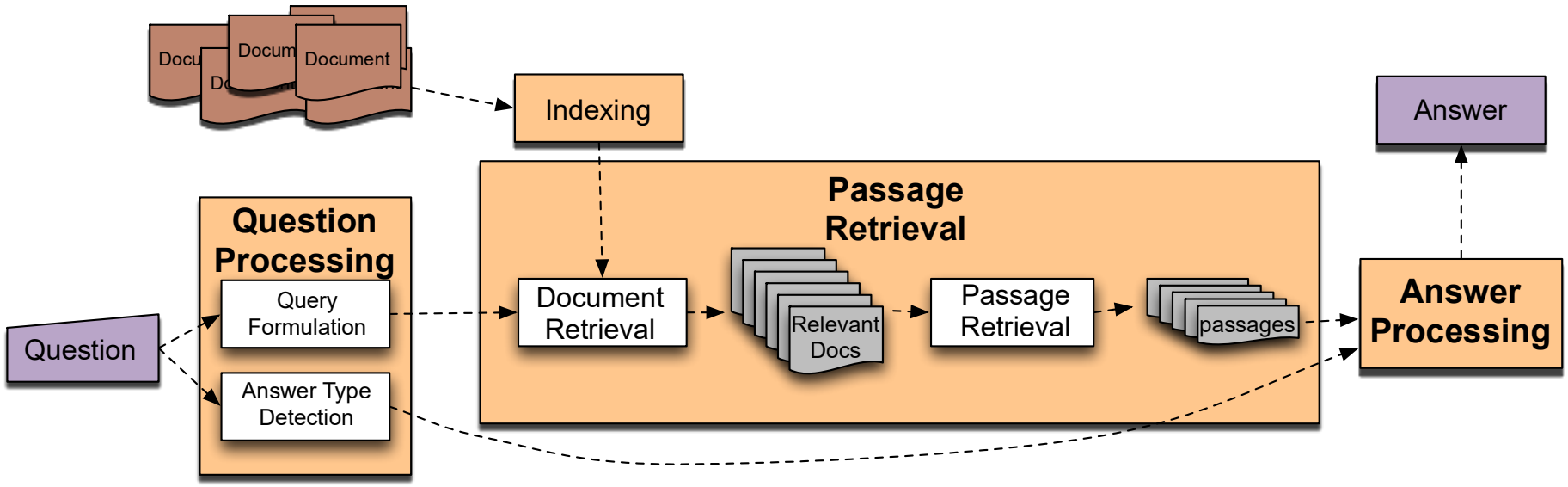
IR-based Question Answering

- Many questions can already be answered by web search

IR-based Question Answering

- But we can do better

IR-based Factoid QA



IR-based Factoid QA

- QUESTION PROCESSING
  - Detect question type, answer type, focus, relations
  - Formulate queries to send to a search engine
- PASSAGE RETRIEVAL
  - Retrieve ranked documents
  - Break into suitable passages and re-rank
- ANSWER PROCESSING
  - Extract candidate answers
  - Rank candidates
    - using evidence from the text and external sources

Knowledge-based approaches (Siri)

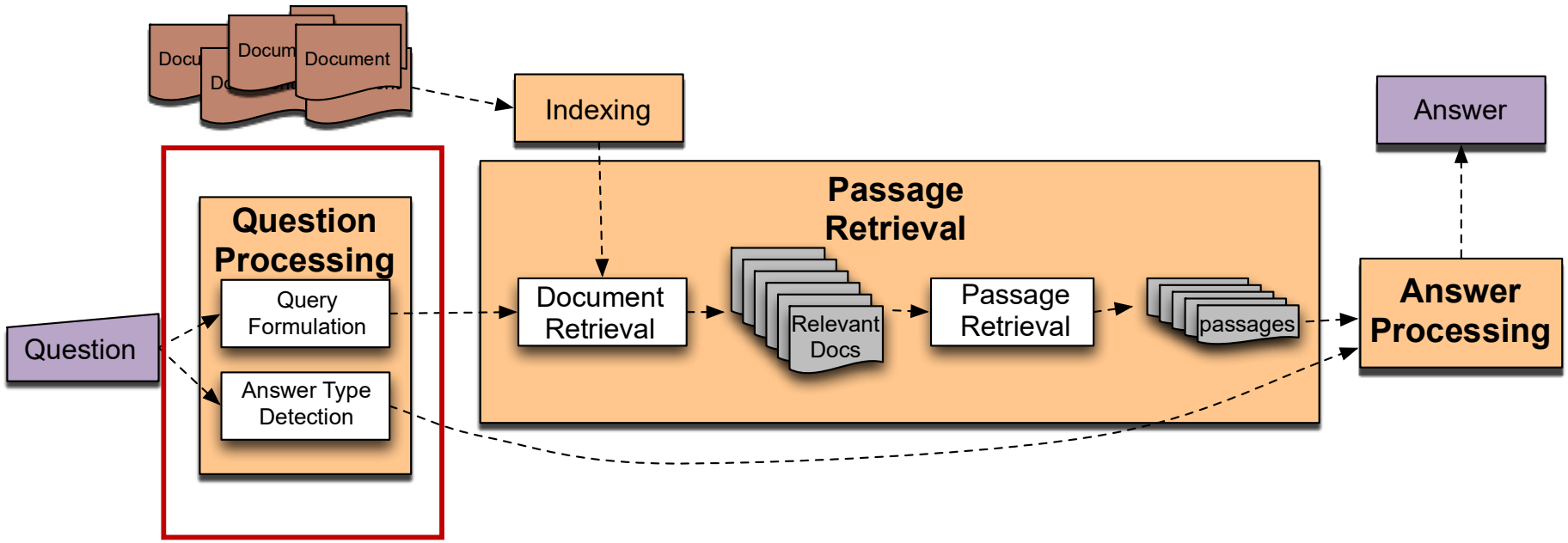
- Build a semantic representation of the query
  - Times, dates, locations, entities, numeric quantities
- Map from this semantics to query structured data or resources
  - Geospatial databases
  - Ontologies (Wikipedia INFOBoxes, DBPedia, WordNet, Yago)
  - Restaurant review sources and reservation services
  - Scientific databases

Hybrid approaches (IBM Watson)

- Build a shallow semantic representation of the query
- Generate answer candidates using IR methods
  - Augmented with ontologies and semi-structured data
- Score each candidate using richer knowledge sources
  - Geospatial databases
  - Temporal reasoning
  - Taxonomical classification

Answer Types and Query Formulation

Factoid Q/A


Docu
Docum
Document
Indexing
Answer
Question Processing
Query Formulation
Answer Type Detection
Passage Retrieval
Document Retrieval
Relevant Docs
Passage Retrieval
passages
Question
Answer Processing

Question Processing

- Answer Type Detection
  - Decide the **named entity type** (person, place) of the answer
- Query Formulation
  - Choose **query keywords** for the IR system
- Question Type classification
  - Is this a definition question, a math question, a list question?
- Focus Detection
  - Find the question words that are replaced by the answer
- Relation Extraction
  - Find relations between entities in the question

Question Processing

They’re the two states you could be reentering if you’re crossing Florida’s northern border

- Answer Type: US state
- Query: two states, border, Florida, north
- Focus: the two states
- Relations: borders(Florida, ?x, north)

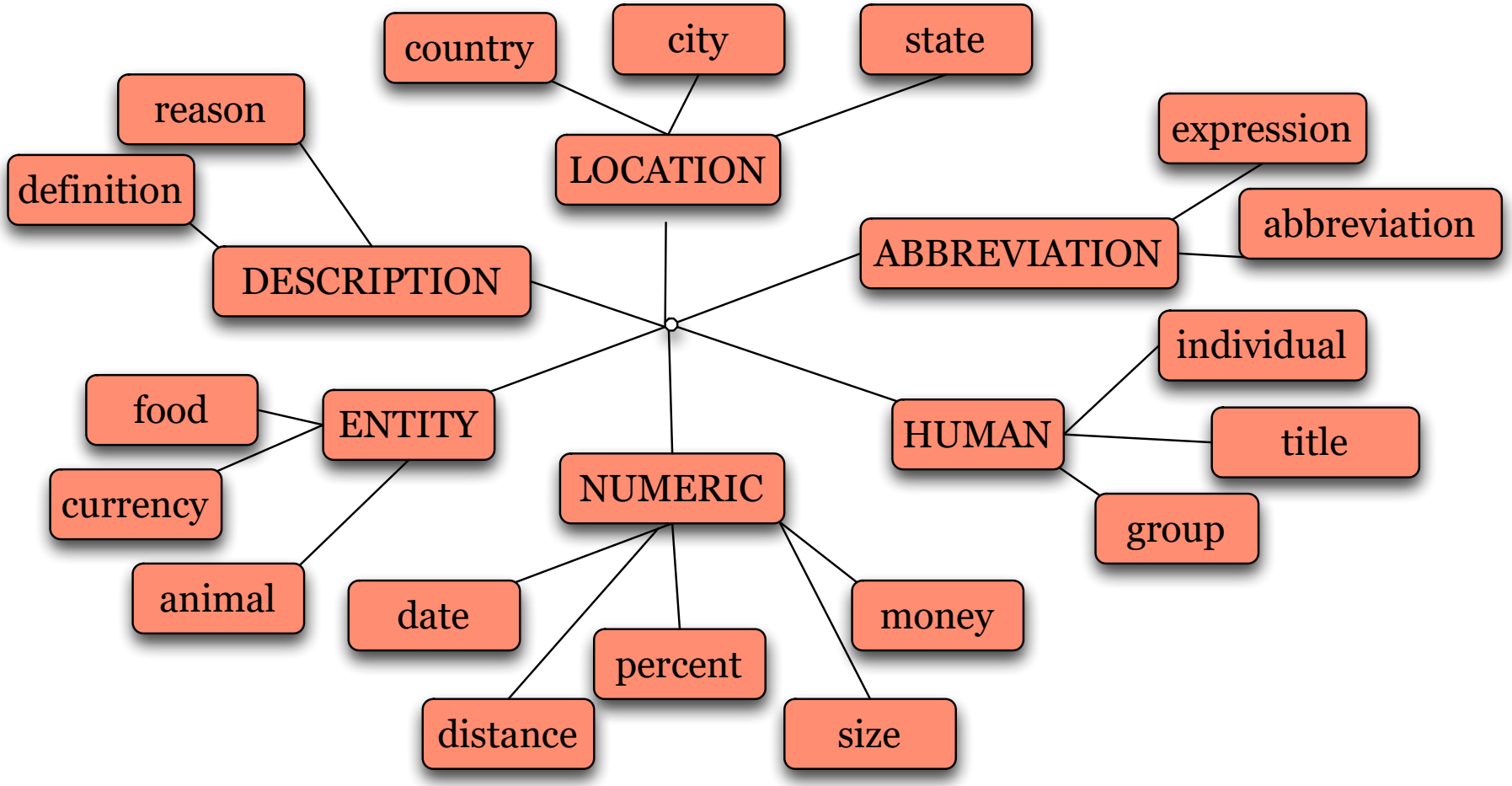
Answer Type Detection: Named Entities

- *Who founded Virgin Airlines?*
  - PERSON
- *What Canadian city has the largest population?*
  - *CITY.*

Answer Type Taxonomy

- 6 coarse classes
  - ABBEVIATION, ENTITY, DESCRIPTION, HUMAN, LOCATION, NUMERIC
- 50 finer classes
  - LOCATION: city, country, mountain…
  - HUMAN: group, individual, title, description
  - ENTITY: animal, body, color, currency…

Part of Li & Roth's Answer Type Taxonomy



Answer Types

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?

Answer types in Jeopardy

- 2500 answer types in 20,000 Jeopardy question sample
- The most frequent 200 answer types cover < 50% of data
- The 40 most frequent Jeopardy answer types

  he, country, city, man, film, state, she, author, group, here, company, president, capital, star, novel, character, woman, river, island, king, song, part, series, sport, singer, actor, play, team, show, actress, animal, presidential, composer, musical, nation, book, title, leader, game

Answer Type Detection

- Hand-written rules
- Machine Learning
- Hybrids

Answer Type Detection

- Regular expression-based rules can get some cases:
  - Who {is|was|are|were} PERSON
  - PERSON (YEAR – YEAR)
- Other rules use the **question headword**:

  (the headword of the first noun phrase after the wh-word)

  - Which **city** in China has the largest number of foreign financial companies?
  - What is the city **flower** of Shenzhen?

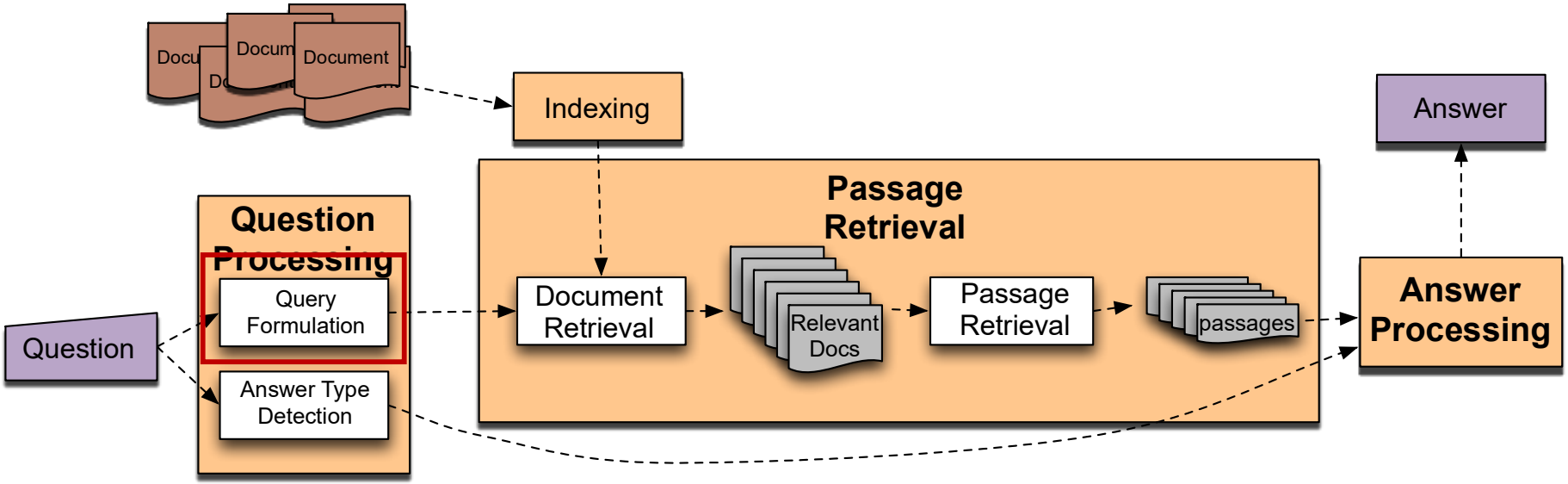
Answer Type Detection

- Most often, we treat the problem as machine learning classification
  - **Define** a taxonomy of question types
  - **Annotate** training data for each question type
  - **Train** classifiers for each question class using a rich set of features.
    - features include those hand-written rules!

Answer Type Detection

- Question words and phrases
- Part-of-speech tags
- Parse features (headwords)
- Named Entities
- Semantically related words

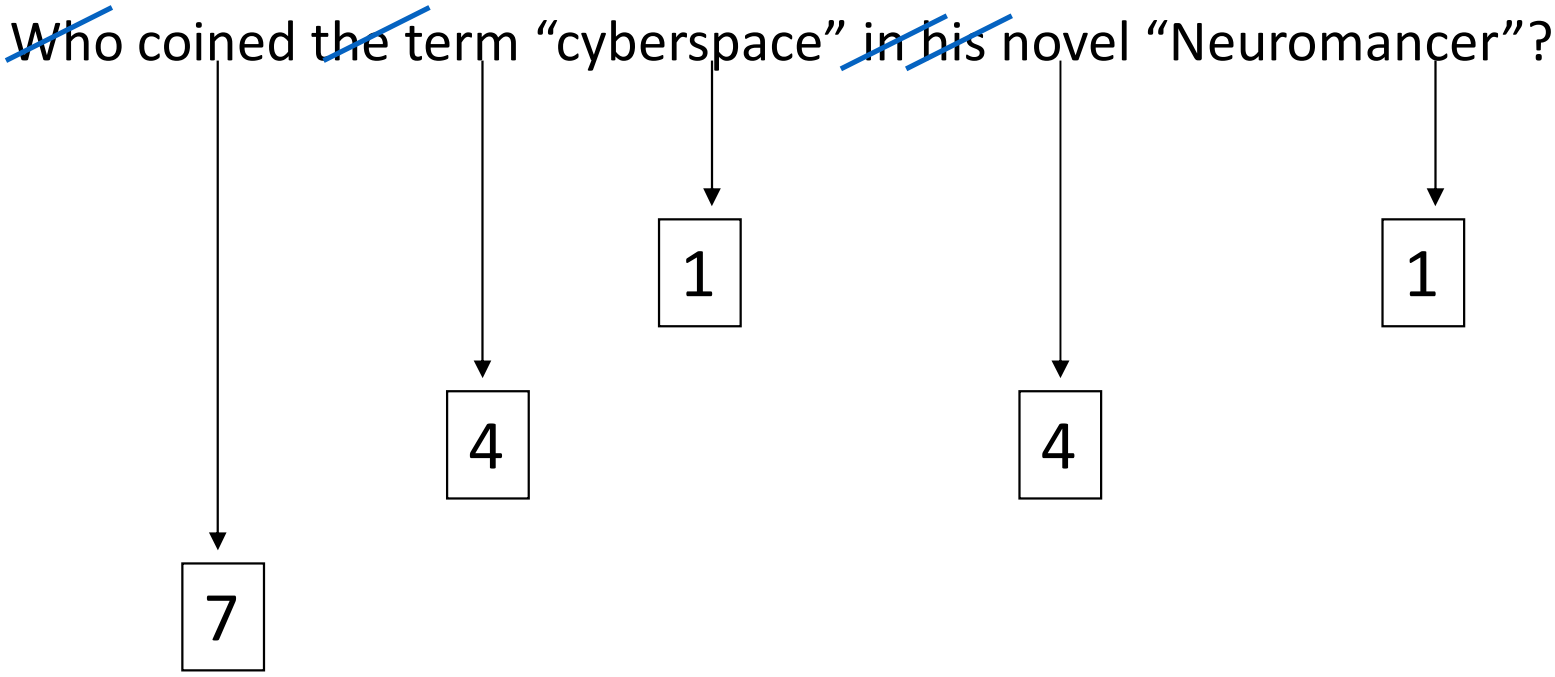
Factoid Q/A


Docu
Docum
Document
Indexing
Answer
Question Processing
Passage Retrieval
Query Formulation
Document Retrieval
Relevant Docs
Passage Retrieval
passages
Answer Processing
Question
Answer Type Detection

Keyword Selection Algorithm

1. Select all non-stop words in quotations
2. Select all NNP words in recognized named entities
3. Select all complex nominals with their adjectival modifiers
4. Select all other complex nominals
5. Select all nouns with their adjectival modifiers
6. Select all other nouns
7. Select all verbs
8. Select all adverbs
9. Select the QFW word (skipped in all previous steps)
10. Select all other words

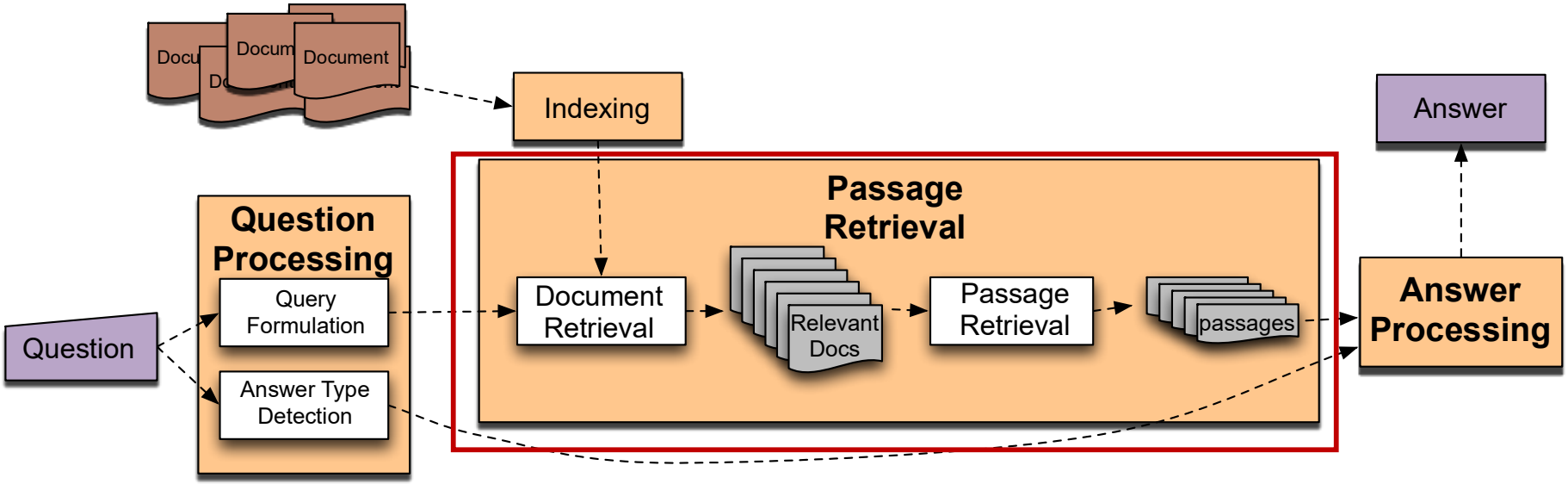
Choosing keywords from the query



cyberspace/1 Neuromancer/1 term/4 novel/4 coined/7

Passage Retrieval and Answer Extraction

Factoid Q/A


Docu
Docum
Document
Indexing
Answer
Passage Retrieval
Question Processing
Query Formulation
Answer Type Detection
Question
Document Retrieval
Relevant Docs
Passage Retrieval
passages
Answer Processing

Passage Retrieval

- Step 1: IR engine retrieves documents using query terms
- Step 2: Segment the documents into shorter units
  - something like paragraphs
- Step 3: Passage ranking
  - Use answer type to help re-rank passages

Documents as vectors

- Now we have a |V|-dimensional vector
- Documents are points or vectors in this space
- very high dimensional: tens of millions of dimensions when you apply this to a web search engine
- These are very sparse vectors-most entries are zero

Query as vectors

- Key idea 1:

  - Do the same for quries: represent them as vectors in the space

- Key idea 2:

  - Rank documents according to their proximity to the query in this space

- Proximity=Similarity of vectors

Formalizing vector space proximity

- First cut: distance between two points
- Euclidean distance?
- Euclidean is a bad idea …
- … because Euclidean distance is large for vectors of different lengths

Formalizing vector space proximity

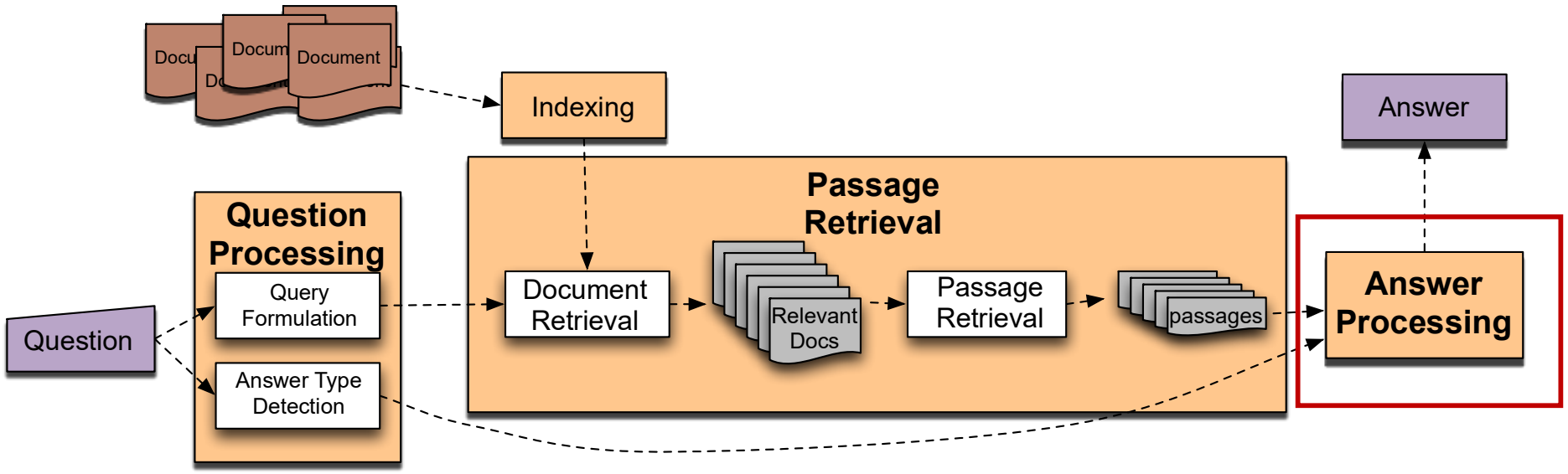
- Take a document d and append it to itself. Call this document d'.
- Semantically d and d' have the content
- The Euclidean distance between the two documents can be quiet large
- The angle between the two documents is 0, corresponding to maximal similarity

- Key idea: Rank documents according to angle with query.

Features for Passage Ranking

- Number of Named Entities of the right type in passage
- Number of query words in passage
- Number of question N-grams also in passage
- Proximity of query keywords to each other in passage
- Longest sequence of question words
- Rank of the document containing passage

Factoid Q/A


Docu
Docum
Document
Indexing
Answer
Passage
Retrieval
Question
Processing
Query
Formulation
Answer Type
Detection
Question
Document
Retrieval
Relevant
Docs
Passage
Retrieval
passages
Answer
Processing

Answer Extraction

- Run an answer-type named-entity tagger on the passages
  - Each answer type requires a named-entity tagger that detects it
  - If answer type is CITY, tagger has to tag CITY
    - Can be full NER, simple regular expressions, or hybrid
- Return the string with the right type:
  - Who is the prime minister of India (PERSON)

Manmohan Singh, Prime Minister of India, had told left leaders that the deal would not be renegotiated.

  - How tall is Mt. Everest? (LENGTH)

The official height of Mount Everest is **29035 feet**

Ranking Candidate Answers

- But what if there are multiple candidate answers!

Q: Who was Queen Victoria's second son?

- Answer Type: **Person**
- Passage:

The Marie biscuit is named after Marie Alexandrovna, the daughter of Czar Alexander II of Russia and wife of Alfred, the second son of Queen Victoria and Prince Albert

Ranking Candidate Answers

- But what if there are multiple candidate answers!

Q: Who was Queen Victoria's second son?

- Answer Type: **Person**
- Passage:

The Marie biscuit is named after **Marie Alexandrovna**, the daughter of **Czar Alexander II of Russia** and wife of **Alfred**, the second son of **Queen Victoria** and **Prince Albert**

Use machine learning: Features for ranking candidate answers

Answer type match: Candidate contains a phrase with the correct answer type.

Pattern match: Regular expression pattern matches the candidate.

Question keywords: number of question keywords in the candidate.

Keyword distance: Distance in words between the candidate and query keywords

Novelty factor: A word in the candidate is not in the query.

Apposition features: The candidate is an appositive to question terms

Punctuation location: The candidate is immediately followed by a comma, period, quotation marks, semicolon, or exclamation mark.

Sequences of question terms: The length of the longest sequence of question terms that occurs in the candidate answer.

Candidate Answer scoring in IBM Watson

- Each candidate answer gets scores from >50 components
  - (from unstructured text, semi-structured text, triple stores)
  - logical form (parse) match between question and candidate
  - passage source reliability
  - geospatial location
    - California is ”southwest of Montana”
  - temporal relationships
  - taxonomic classification

Common Evaluation Metrics

Accuracy (does answer match gold-labeled answer?)

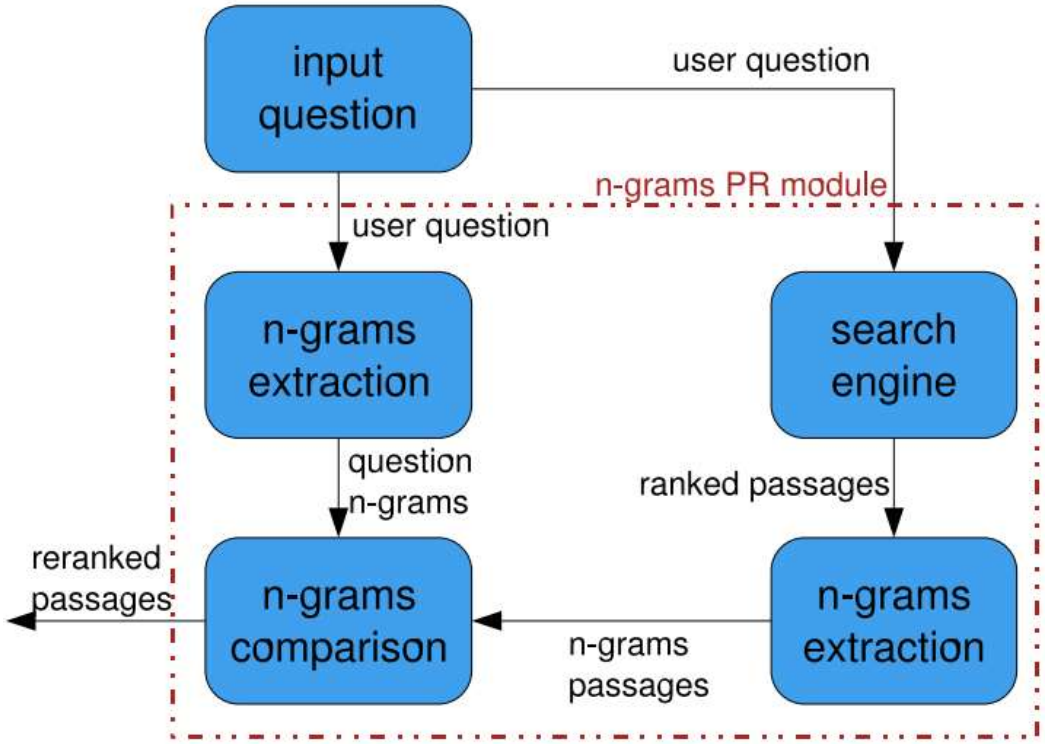
Mean Reciprocal Rank

- For each query return a ranked list of M candidate answers.
- Query score is 1/Rank of the first correct answer
  - *If first answer is correct: 1*
  - *else if second answer is correct: ½*
  - *else if third answer is correct: ⅓, etc.*
  - *Score is 0 if none of the M answers are correct*
- Take the mean over all N queries

$$MRR = \frac{\sum_{i=1}^{N} \frac{1}{rank_i}}{N}$$

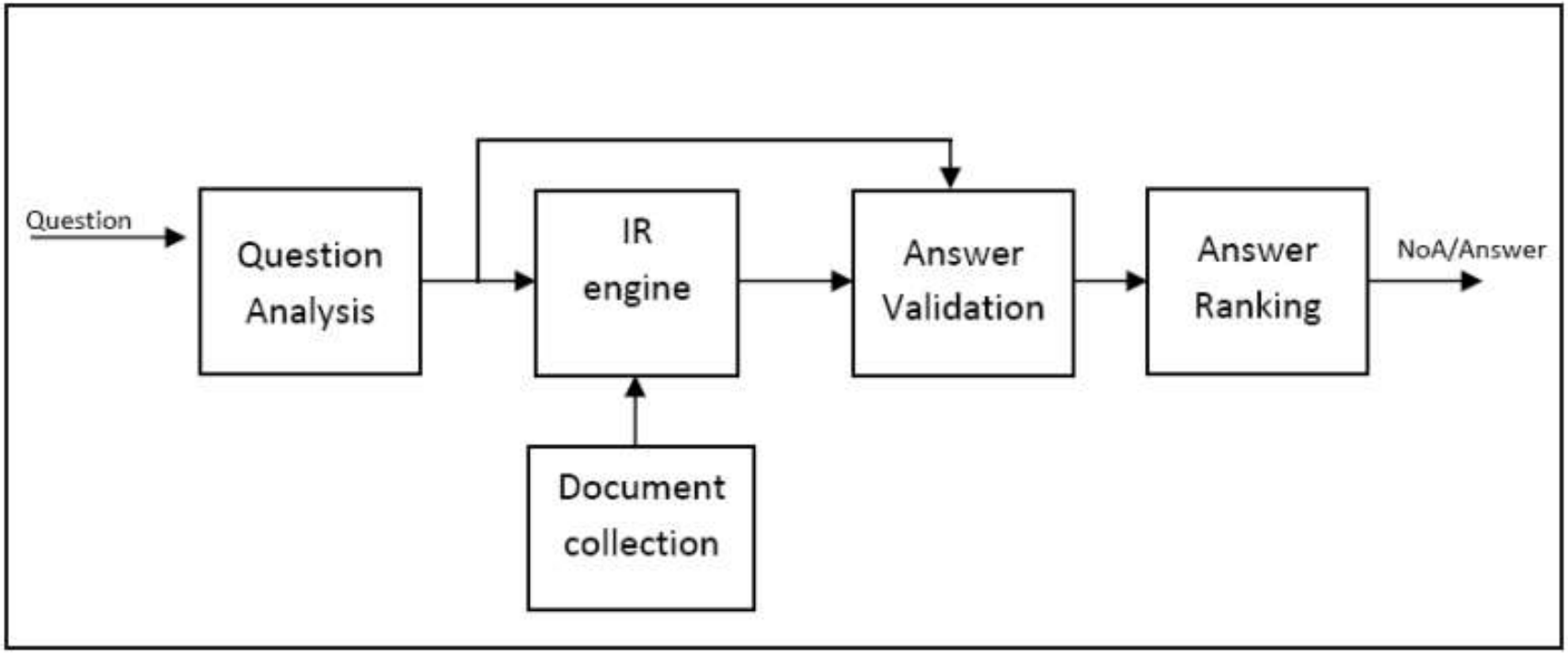
A passage retrieval system for question answering (2005)

- The ranking of passages obtained by the PR system is rearranged
- The new ranking is based on finding the n-gram structures of the question



A QA System based on Information Retrieval and Validation (2010)

- based on applying an Information Retrieval engine of high performance and a validation step for removing incorrect answers.



Using Knowledge in QA

Relation Extraction

- Answers: Databases of Relations
  - born-in(“Emma Goldman”, “June 27 1869”)
  - author-of(“Cao Xue Qin”, “Dream of the Red Chamber”)
  - Draw from Wikipedia infoboxes, DBpedia, FreeBase, etc.
- Questions: Extracting Relations in Questions

  Whose granddaughter starred in E.T.?

```
(acted-in ?x “E.T.”)
  (granddaughter-of ?x ?y)
```

Temporal Reasoning

- Relation databases
  - (biographical dictionaries, etc.)
- IBM Watson

  ”In 1594 he took a job as a tax collector in Andalusia”

  Candidates:
    - Thoreau is a bad answer (born in 1817)
    - Cervantes is possible (was alive in 1594)

Geospatial knowledge (containment, directionality, borders)

- Beijing is a good answer for "Asian city"
- California is "southwest of Montana"
- geonames.org:

www.geonames.org/search.html?q=palo+alto&country=

GeoNames Home | Postal Codes | Download / Webservice | About — login

palo alto | all countries | search | show on map | [advanced search]

459 records found for "palo alto"

	Name	Country	Feature class	Latitude	Longitude
1	Palo Alto Palo Al'to,Palo Alto,pa luo ao duo,paroaruto,Пало Алто,Пало Альто,פאלו אלטו,パロアルト,帕羅奧多	United States, California Santa Clara County	populated place population 64,403, elevation 9m	N 37° 26' 30''	W 122° 8' 34''
2	Palo Alto Township Palo Alto Township	United States, Iowa Jasper County	administrative division elevation 256m	N 41° 38' 15''	W 93° 2' 57''
3	Borough of Palo Alto	United States, Pennsylvania Schuylkill County	administrative division population 1,032, elevation 210m	N 40° 41' 21''	W 76° 10' 2''

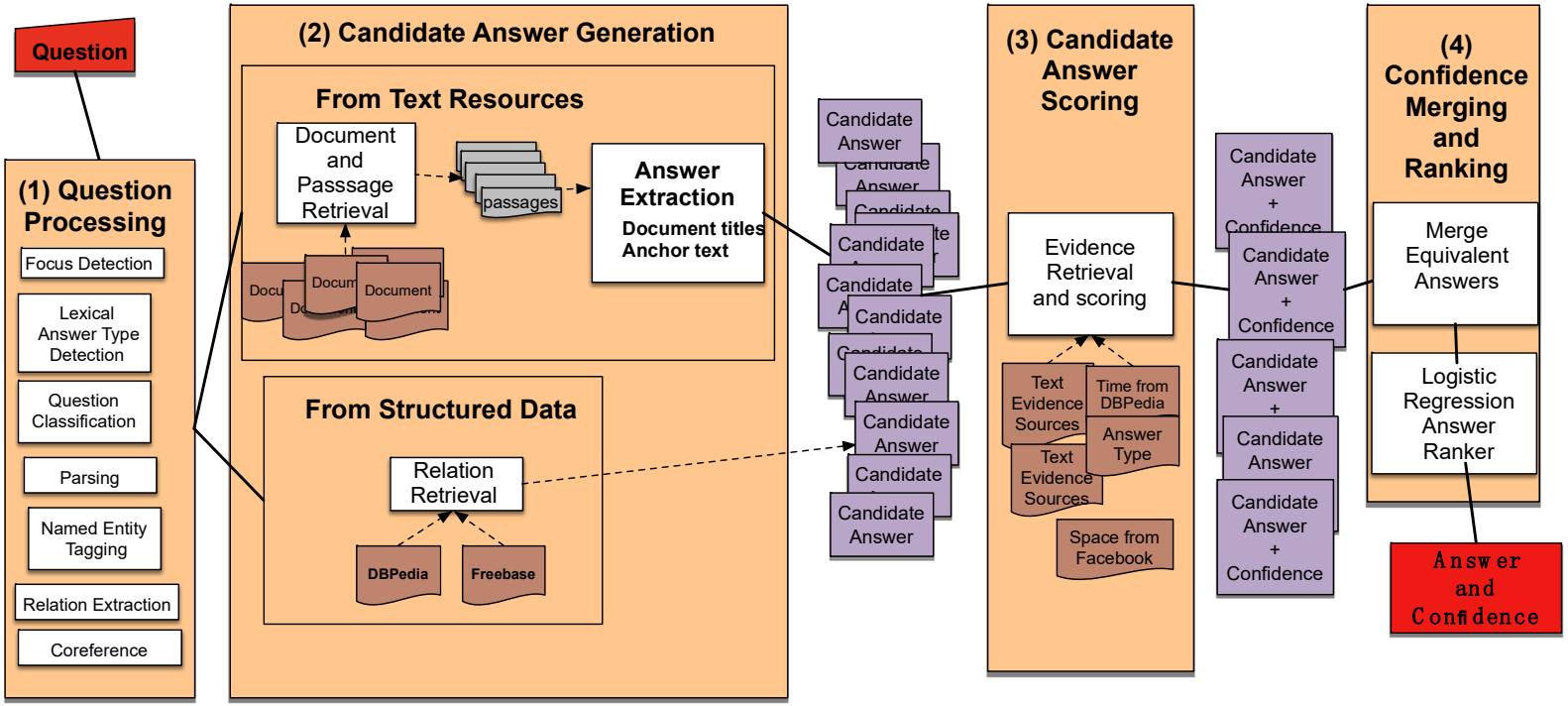
Context and Conversation in Virtual Assistants like Siri

- Coreference helps resolve ambiguities
  - U: “Book a meeting room at T6 at 7:00 with **Mr. Xu**”
  - U: “Also send **him** an email reminder”

- Clarification questions:
  - U: “Wuxi xiaolongbao”
  - S: “Did you mean xiaolongbao restaurants in Wuxi or Wuxi-style xiaolongbao?”

Question Answering in Watson (Deep QA)

The Architecture of Watson



Stage 1: Question Processing

- Parsing
- Named Entity Tagging
- Relation Extraction
- Focus
- Answer Type
- Question Classification

Stage 1: Question Processing

Poets and Poetry: **He** was a bank clerk in the **Yukon** before he published "**Songs of a Sourdough**" in 1907.

THEATRE: A new play based on **this Sir Arthur Conan Doyle** canine classic opened on the London stage in 2007.

- Relations:

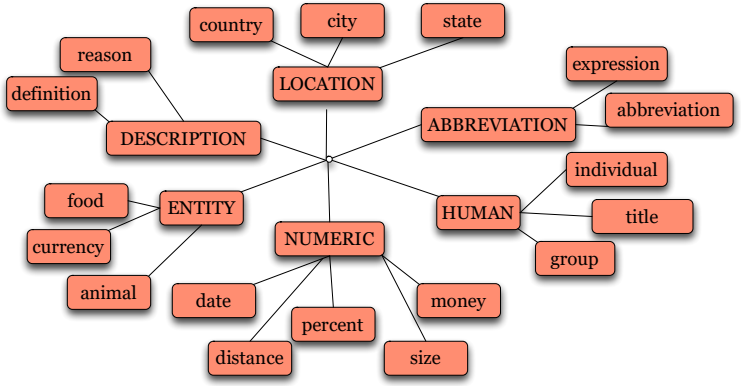
  ("he", author-of, "songs of a sourdough")
  ("he", was, "bank clerk")
  ("bank clerk", in, "yukon")

Focus extraction

- **Focus**: the part of the question that co-refers with the answer
- Replace it with answer to find a supporting passage.
- Extracted by hand-written rules
  - "Extract any noun phrase with determiner this”
  - “Extracting pronouns *she, he, hers, him, “*

Answer Type

- The semantic class of the answer
- But for Jeopardy the TREC answer type taxonomy is insufficient
- Watson team investigated 20,000 questions
  - 100 named entities only covered <50% of the questions!
- Instead: Extract lots of words: 5,000 for those 20,000 questions



Answer Type

- Answer types extracted by hand-written rules
  - Syntactic headword of the focus.
  - Words that are coreferent with the focus
  - Jeopardy category, if refers to compatible entity.

Poets *and Poetry:* ***He*** *was a bank* ***clerk*** *in the Yukon before he published "Songs of a Sourdough" in 1907.*

Stage 2: Candidate Answer Generation

- combine the processed question with external documents and other knowledge sources to suggest many candidate answers
  - from both text documents
  - from structured knowledge bases

Extracting candidate answers from triple stores

- If we extracted a relation from the question

  … he published “Songs of a sourdough”

  ```
  (author-of ?x “Songs of a sourdough”)
  ```

- We just query a triple store
  - Wikipedia infoboxes, DBpedia, FreeBase, etc.
  - born-in (“Emma Goldman”, “June 27 1869”)
  - author-of (“Cao Xue Qin”, “Dream of the Red Chamber”)
  - author-of (“Songs of a sourdough”, “Robert Service”)

Extracting candidate answers from text: get documents/passages

- Do standard IR-based QA to get documents

- Robert Redford and Paul Newman starred in this depression-era grifter flick.
  - (2.0 Robert Redford)
  - (2.0 Paul Newman)
  - star depression era grifter
  - (1.5 flick)

Stage 3: Candidate Answer Scoring

- Use lots of sources of evidence to score an answer
  - more than 50 scorers
- **Answer type** is a big one
  - Different in Watson than in pure IR factoid QA
  - In pure IR factoid QA, answer type is used to strictly filter answers
  - In Watson, answer type is just one of many pieces of evidence

Answer Type for Scoring Candidates

- Given:
  - candidate answer & answer type
- Return a score: can answer type be a subclass of this answer type?
- Candidate: "*difficulty swallowing*" & manifestation
  1. Check DBPedia, WordNet, etc
     - *difficulty swallowing* -> (DBPedia) *Dysphagia* -> (WordNet) *Dysphagia*
     - manifestation -> (WordNet) *Condition*
  2. Check if "Dysphagia" is a "Condition" in WordNet
     - Wordnet for dysphagia

Relations for scoring

- **Q**: This hockey defenseman ended his career on June 5, 2008
- **Passage**: On June 5, 2008, Wesley announced his retirement after his 20th NHL season

- Question and passage have very few keywords in common
- But both have the Dbpedia relation `ActiveYearsEndDate()`

Temporal Reasoning for Scoring Candidates

- Relation databases
  - (and obituaries, biographical dictionaries, etc.)
- IBM Watson

  ”In 1594 he took a job as a tax collector in Andalusia”

  Candidates:
    - Thoreau is a bad answer (born in 1817)
    - Cervantes is possible (was alive in 1594)

Geospatial knowledge (containment, directionality, borders)

- Beijing is a good answer for "Asian city"
- California is "southwest of Montana"
- geonames.org:

Text-retrieval-based answer scorer

- Generate a query from the question and retrieve passages
- Replace the focus in the question with the candidate answer
  - See how well it fits the passages.
- Robert Redford and Paul Newman starred in **this depression-era grifter flick**
- Robert Redford and Paul Newman starred in The Sting

Robert Redford - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Robert_Redford** ▾ Wikipedia ▾
Redford starred in Sydney Pollack's Out of Africa (1985), which was an .... by William Goldman, in which he was paired for the first time with **Paul Newman**. ... the blockbuster crime caper The **Sting** (1973), which became one of the top 20 ...

Stage 4: Answer Merging and Scoring

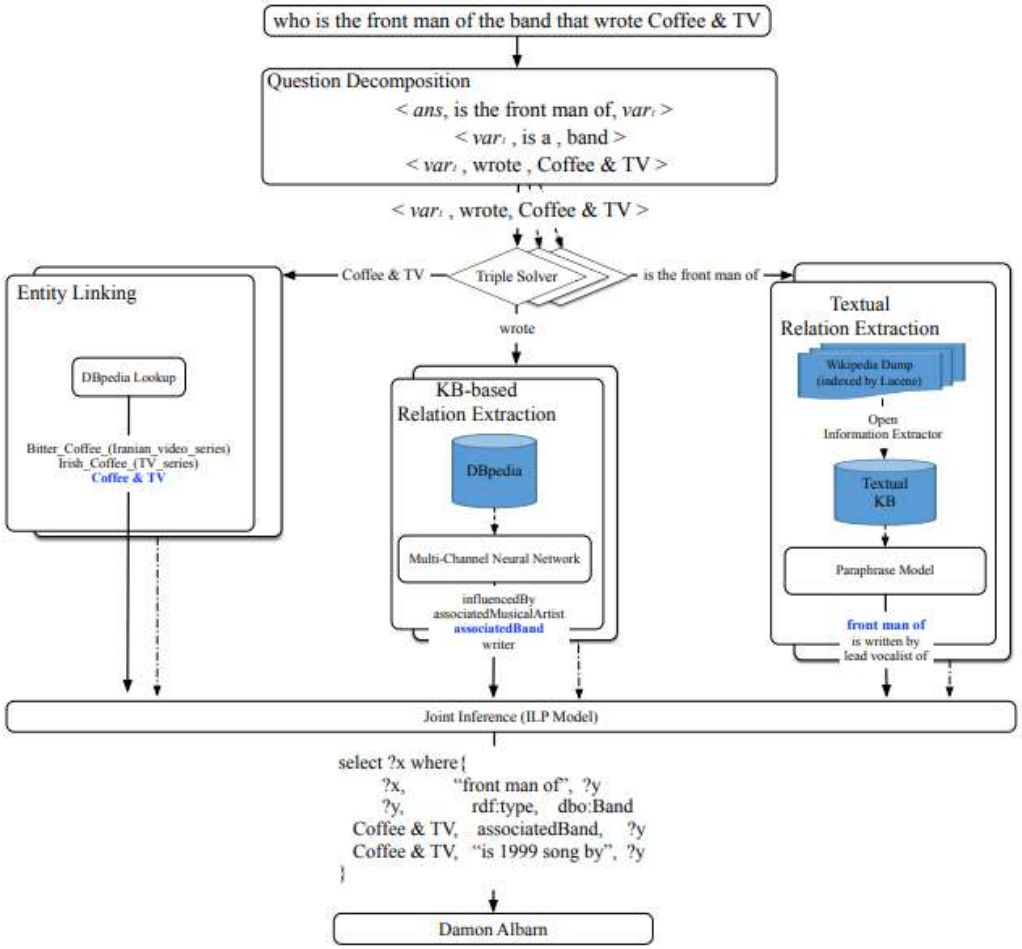
- Now we have a list candidate answers each with a score vector
  - `J.F.K [.5 .4 1.2 33 .35 … ]`
  - `John F. Kennedy [.2 .56 5.3 2 …]`
- Merge equivalent answers*: J.F.K.* and *John F. Kennedy*
  - Use Wikipedia dictionaries that list synonyms:
    - J*FK, John F. Kennedy, John Fitzgerald Kennedy, Senator John F. Kennedy, President Kennedy, Jack Kennedy*
  - Use stemming and other morphology

Stage 4: Answer Merging and Scoring

- Build a classifier to take answers and a score vector and assign a probability
- Train on datasets of hand-labeled correct and incorrect answers.

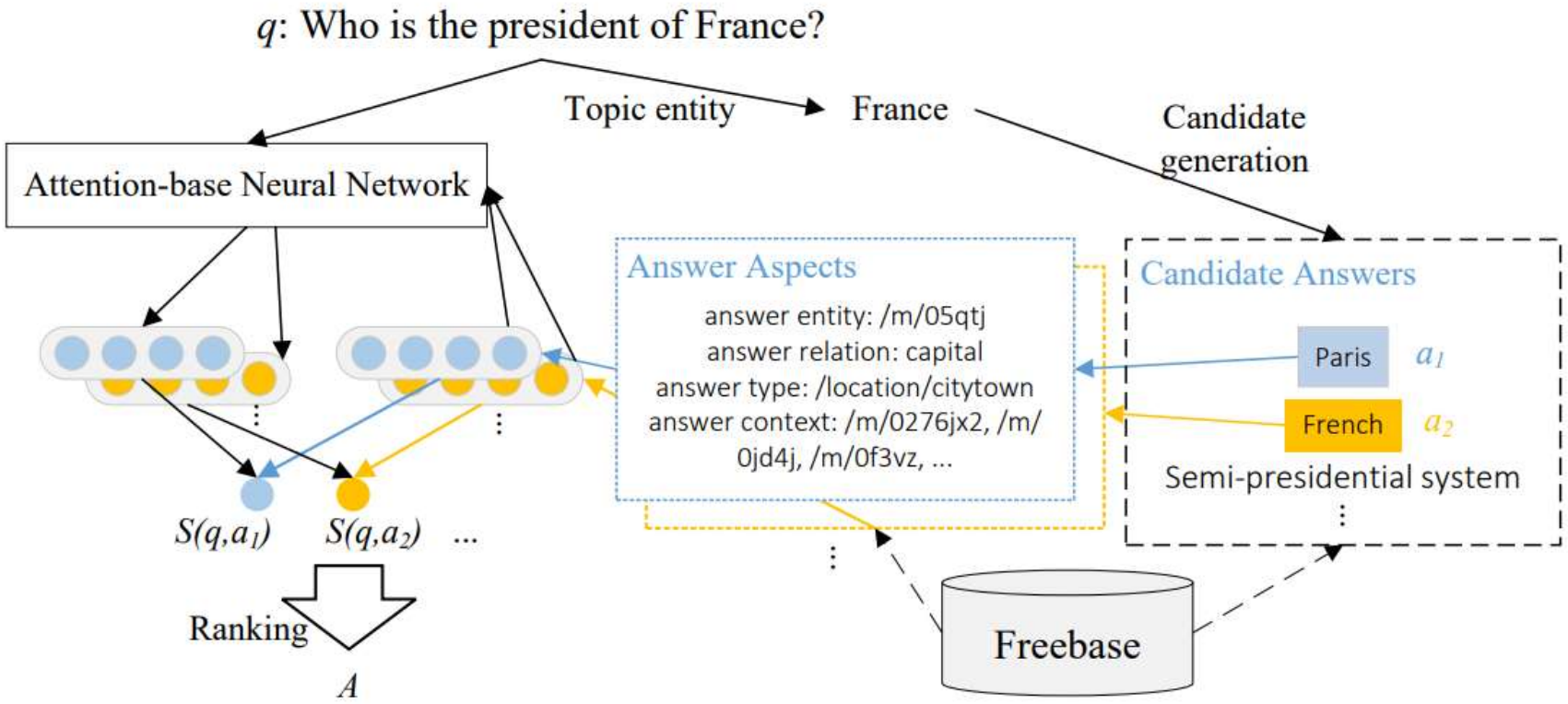
Hybrid Question Answering over Knowledge Base and Free Text (ACL 2016)

- perform the local predictions
- further infer on the retained candidate entities



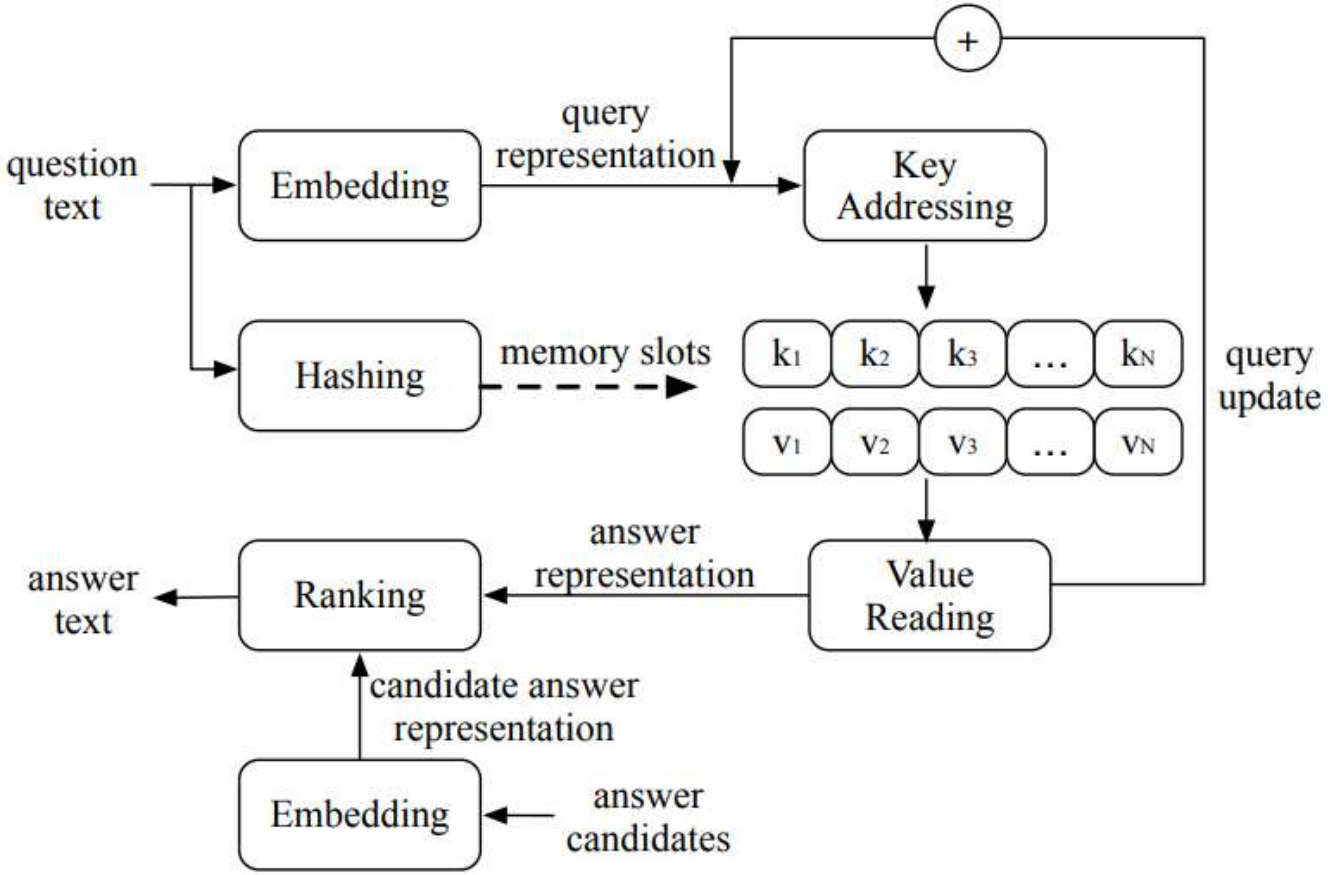
Question answering over KB with global knowledge information (arxiv 2016)

- a neural attention-based model to represent the questions dynamically according to the different focuses of various candidate answer aspects.



Memory neural networks for knowledge based question answering (ACL 2019)

- a novel mechanism to perform interpretable reasoning for complex questions.



Machine Reading Comprehension

Motivation

- With massive collections of full-text documents, i.e., the web , simply returning relevant documents is of limited use
- Rather, we often want answers to our questions
  - Especially on mobile
  - Or using a digital assistant device, like Alexa, Google Assistant, …
- We can factor this into two parts:
  - 1. Finding documents that (might) contain an answer
    - Which can be handled by traditional information retrieval/web search
  - 2. Finding an answer in a paragraph or a document
    - This problem is often termed Reading Comprehension

A Brief History of Reading Comprehension

- Much early NLP work attempted reading comprehension
  - Schank, Abelson, Lehnert et al. c. 1977 – "Yale A.I. Project"
- Revived by Lynette Hirschman in 1999:
  - Could NLP systems answer human reading comprehension questions for 3rd to 6th graders? Simple methods attempted.
- Revived again by Chris Burges in 2013 with MCTest
  - Again answering questions over simple story texts
- Floodgates opened in 2015/16 with the production of large datasets which permit supervised neural systems to be built
  - Hermann et al. (NIPS 2015) DeepMind CNN/DM dataset
  - Rajpurkar et al. (EMNLP 2016) SQuAD
  - MS MARCO, TriviaQA, RACE, NewsQA, NarrativeQA, …

Machine Comprehension

- “A machine comprehends a passage of text if, for any question regarding that text that can be answered correctly by a majority of native speakers, that machine can provide a string which those speakers would agree both answers that question, and does not contain information irrelevant to that question.”

Towards the Machine Comprehension of Text: An Essay

Christopher J.C. Burges
Microsoft Research
One Microsoft Way
Redmond, WA 98052, USA

December 23, 2013

MCTest Reading Comprehension

Passage (*P*) + Question (*Q*) ⟶ Answer (*A*)

P

Alyssa got to the beach after a long trip. She's from Charlotte. She traveled from Atlanta. She's now in Miami. She went to Miami to visit some friends. But she wanted some time to herself at the beach, so she went there first. After going swimming and laying out, she went to her friend Ellen's house. Ellen greeted Alyssa and they both had some lemonade to drink. Alyssa called her friends Kristin and Rachel to meet at Ellen's house…….

Q Why did Alyssa go to Miami?

A To visit some friends

Stanford Question Answering Dataset (SQuAD)

(Rajpurkar et al., 2016)

Question: Which team won Super Bowl 50?

Passage

Super Bowl 50 was an American football game to determine the champion of the National Football League (NFL) for the 2015 season. The American Football Conference (AFC) champion Denver Broncos defeated the National Football Conference (NFC) champion Carolina Panthers 24–10 to earn their third Super Bowl title. The game was played on February 7, 2016, at Levi's Stadium in the San Francisco Bay Area at Santa Clara, California.

100k examples

Answer must be a span in the passage

A.k.a. extractive question answering

Stanford Question Answering Dataset (SQuAD)

Private schools, also known as independent schools, non-governmental, or nonstate schools, are not administered by local, state or national governments; thus, they retain the right to select their students and are funded in whole or in part by charging their students tuition, rather than relying on mandatory taxation through public (government) funding; at some private schools students may be able to get a scholarship, which makes the cost cheaper, depending on a talent the student may have (e.g. sport scholarship, art scholarship, academic scholarship), financial need, or tax credit scholarships that might be available.

Along with non-governmental and nonstate schools, what is another name for private schools?

Gold answers: ① independent ② independent schools ③ independent schools

Along with sport and art, what is a type of talent scholarship?

Gold answers: ① academic ② academic ③ academic

Rather than taxation, what are private schools largely funded by?

Gold answers: ① tuition ② charging their students tuition ③ tuition

SQuAD evaluation, v1.1

- Authors collected 3 gold answers
- Systems are scored on two metrics:
  - Exact match: 1/0 accuracy on whether you match one of the 3 answers
  - F1: Take system and each gold answer as bag of words, evaluate
    Precision = $\frac{TP}{TP+FP}$, Recall = $\frac{TP}{TP+FN}$, harmonic mean F1 = $\frac{2PR}{P+R}$
    Score is (macro-)average of per-question F1 scores
- F1 measure is seen as more reliable and taken as primary
  - It's less based on choosing exactly the same span that humans chose, which is susceptible to various effects, including line breaks
- Both metrics ignore punctuation and articles (a, an, the only)

SQuAD v1.1 leaderboard, 2016

	EM	F1
Human	82.3	91.2
BiDAF (ensemble)	73.3	81.1
Dynamic Coattention Networks (ensemble)	71.6	80.4
R-net (ensemble)	72.1	79.7
BiDAF (signle model)	68.0	77.3
Multi-Perspective Matching (ensemble)	68.2	77.2

SQuAD v1.1 leaderboard, 2019 – it's solved!

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar et al. '16)	82.304	91.221
1 Oct 05, 2018	BERT (ensemble) *Google AI Language* https://arxiv.org/abs/1810.04805	**87.433**	**93.160**
2 Oct 05, 2018	BERT (single model) *Google AI Language* https://arxiv.org/abs/1810.04805	85.083	91.835
2 Sep 09, 2018	nlnet (ensemble) *Microsoft Research Asia*	85.356	91.202
2 Sep 26, 2018	nlnet (ensemble) *Microsoft Research Asia*	85.954	91.677
3 Jul 11, 2018	QANet (ensemble) *Google Brain & CMU*	84.454	90.490
4 Jul 08, 2018	r-net (ensemble) *Microsoft Research Asia*	84.003	90.147
5 Mar 19, 2018	QANet (ensemble) *Google Brain & CMU*	83.877	89.737
5 Sep 09, 2018	nlnet (single model) *Microsoft Research Asia*	83.468	90.133

SQuAD 2.0

- A defect of SQuAD 1.0 is that all questions have an answer in the paragraph
  - Systems (implicitly) rank candidates and choose the best one
  - You don't have to judge whether a span answers the question
- In SQuAD 2.0, 1/3 of the training questions have no answer, and about 1/2 of the dev/test questions have no answer
  - For NoAnswer examples, NoAnswer receives a score of 1, and any other response gets 0, for both exact match and F1
- Simplest system approach to SQuAD 2.0:
  - Have a threshold score for whether a span answers a question
  - Or you could have a second component that confirms answering
    - Like Natural Language Inference (NLI) or "Answer validation"

SQuAD 2.0 Example

Genghis Khan united the Mongol and Turkic tribes of the steppes and became Great Khan in 1206. He and his successors expanded the Mongol empire across Asia. Under the reign of Genghis' third son, Ögedei Khan, the Mongols destroyed the weakened Jin dynasty in 1234, conquering most of northern China. Ögedei offered his nephew Kublai a position in Xingzhou, Hebei. Kublai was unable to read Chinese but had several Han Chinese teachers attached to him since his early years by his mother Sorghaghtani. He sought the counsel of Chinese Buddhist and Confucian advisers. Möngke Khan succeeded Ögedei's son, Güyük, as Great Khan in 1251. He

When did Genghis Khan kill Great Khan?

Gold Answers: <No Answer>

Prediction: 1234 [from Microsoft nlnet]

SQuAD 2.0 leaderboard, 2019-02-07

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Jan 15, 2019	BERT + MMFT + ADA (ensemble) *Microsoft Research Asia*	**85.082**	**87.615**
2 Jan 10, 2019	BERT + Synthetic Self-Training (ensemble) *Google AI Language* https://github.com/google-research/bert	84.292	86.967
3 Dec 13, 2018	BERT finetune baseline (ensemble) *Anonymous*	83.536	86.096
4 Dec 16, 2018	Lunet + Verifier + BERT (ensemble) *Layer 6 AI NLP Team*	83.469	86.043
4 Dec 21, 2018	PAML+BERT (ensemble model) *PINGAN GammaLab*	83.457	86.122
5 Dec 15, 2018	Lunet + Verifier + BERT (single model) *Layer 6 AI NLP Team*	82.995	86.035

Good systems are great, but still basic NLU errors

The Yuan dynasty is considered both a successor to the Mongol Empire and an imperial Chinese dynasty. It was the khanate ruled by the successors of Möngke Khan after the division of the Mongol Empire. In official Chinese histories, the Yuan dynasty bore the Mandate of Heaven, following the Song dynasty and preceding the Ming dynasty. The dynasty was established by Kublai Khan, yet he placed his grandfather Genghis Khan on the imperial records as the official founder of the

What dynasty came before the Yuan?

Gold Answers: ① Song dynasty ② Mongol Empire
③ the Song dynasty

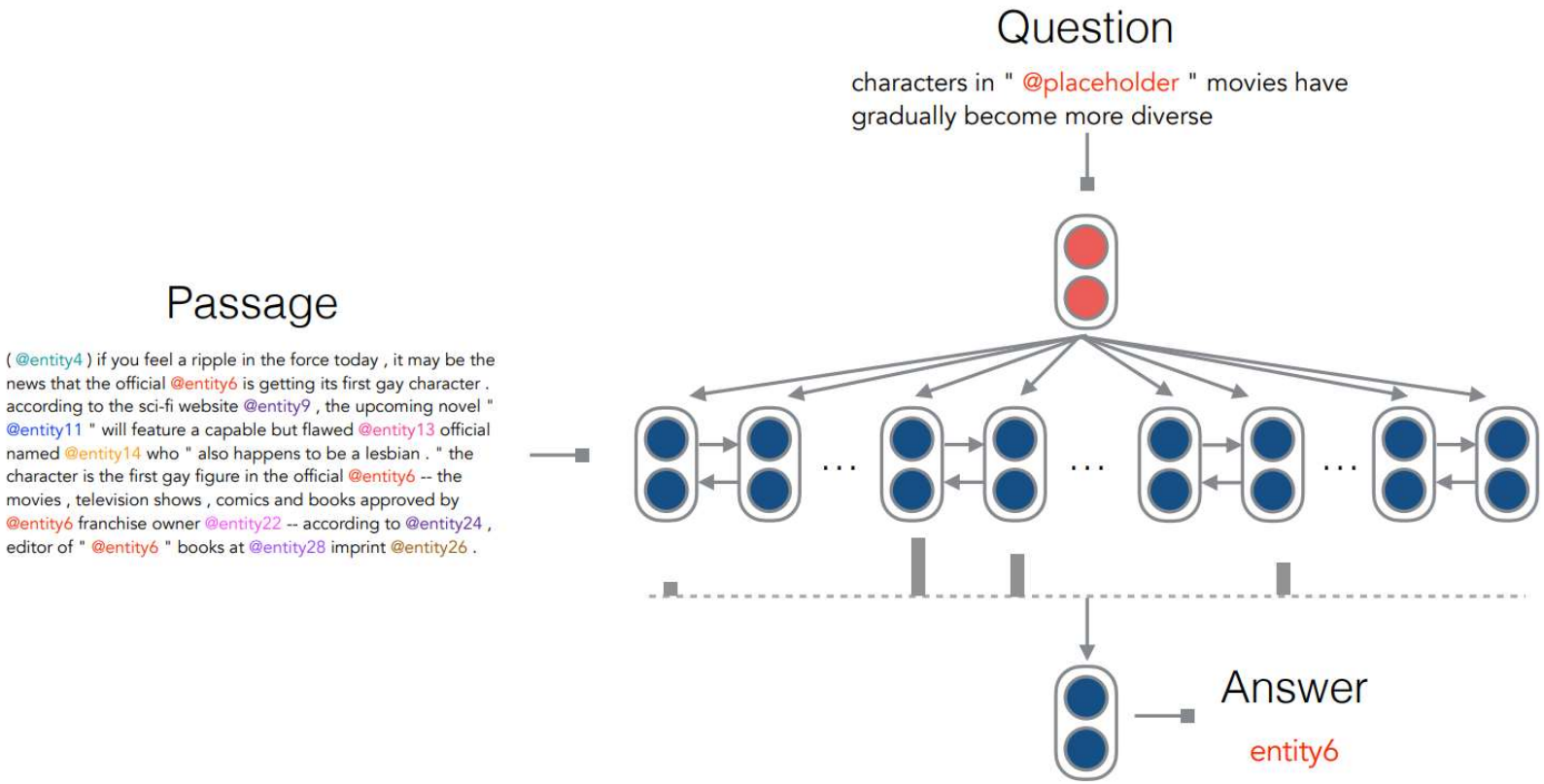
Prediction: Ming dynasty [BERT (single model) (Google AI)]

SQuAD limitations

- SQuAD has a number of other key limitations too:
  - Only span-based answers (no yes/no, counting, implicit why)
  - Questions were constructed looking at the passages
    - Not genuine information needs
    - Generally greater lexical and syntactic matching between questions and answer span than you get IRL
  - Barely any multi-fact/sentence inference beyond coreference
- Nevertheless, it is a well-targeted, well-structured, clean dataset
  - It has been the most used and competed on QA dataset
  - It has also been a useful starting point for building systems in industry (though in-domain data always really helps!)

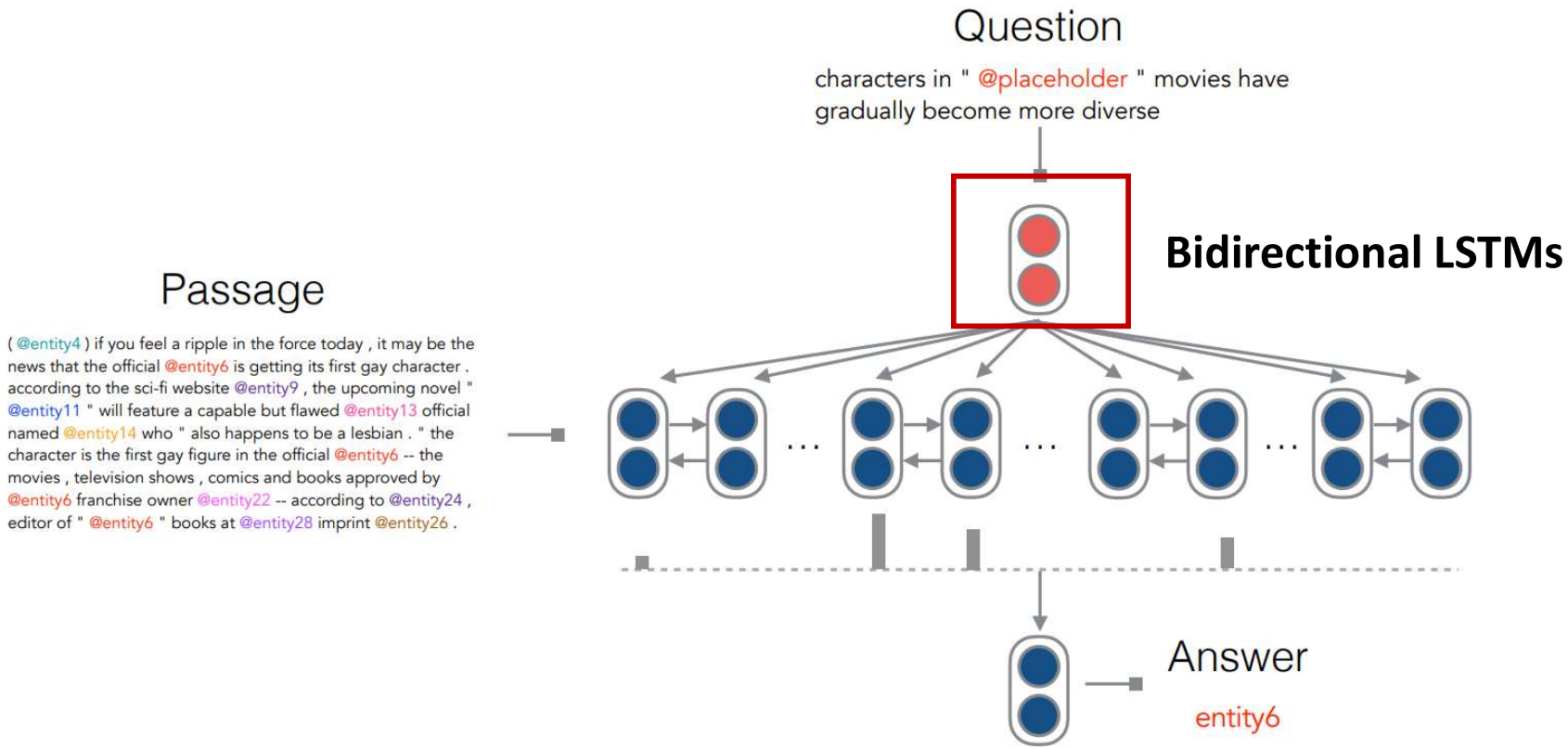
The Attentive Reader

- Demonstrated a minimal, highly successful architecture for reading comprehension and question answering
- Became known as the Attentive Reader



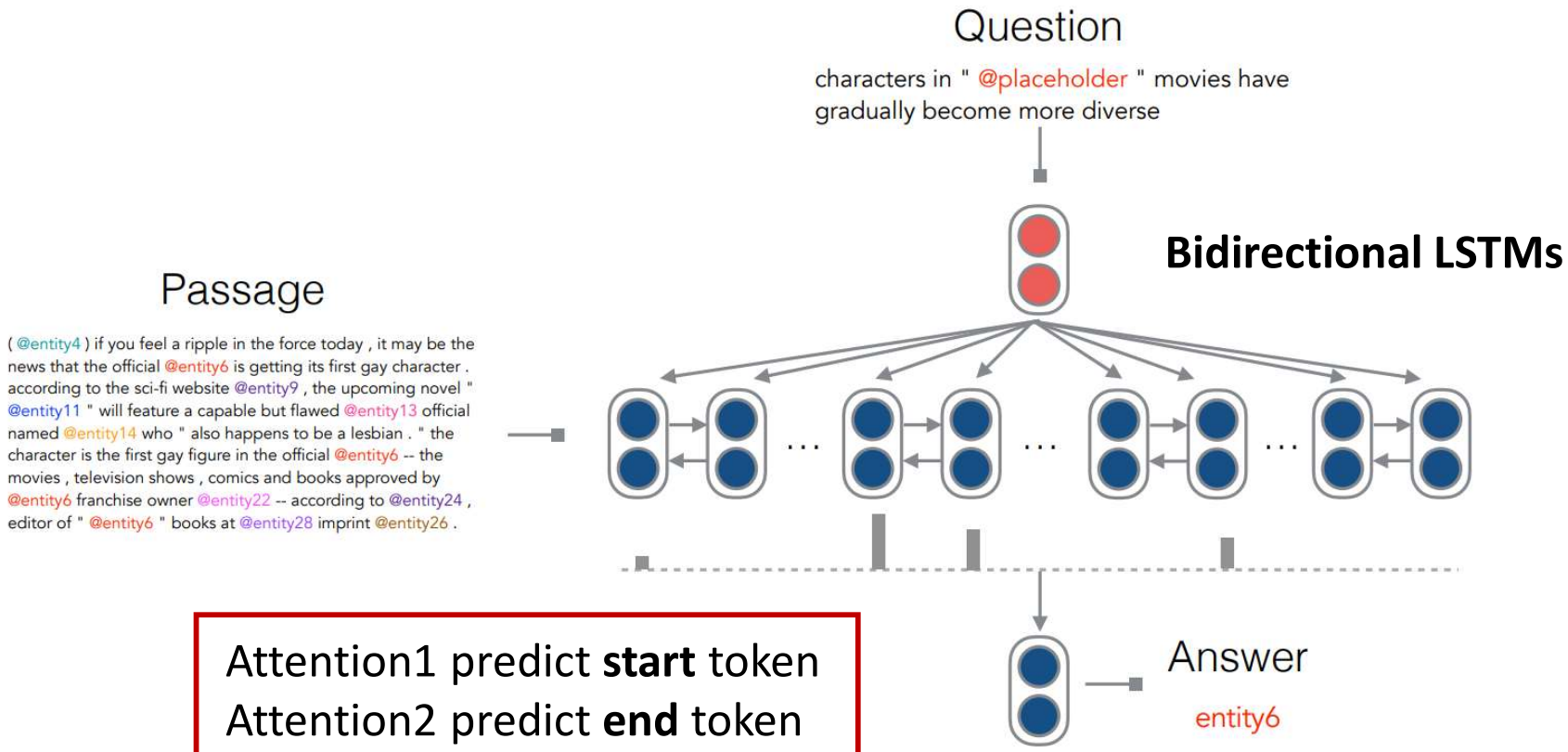
The Attentive Reader

- Demonstrated a minimal, highly successful architecture for reading comprehension and question answering
- Became known as the Attentive Reader



The Attentive Reader

- Demonstrated a minimal, highly successful architecture for reading comprehension and question answering
- Became known as the Attentive Reader

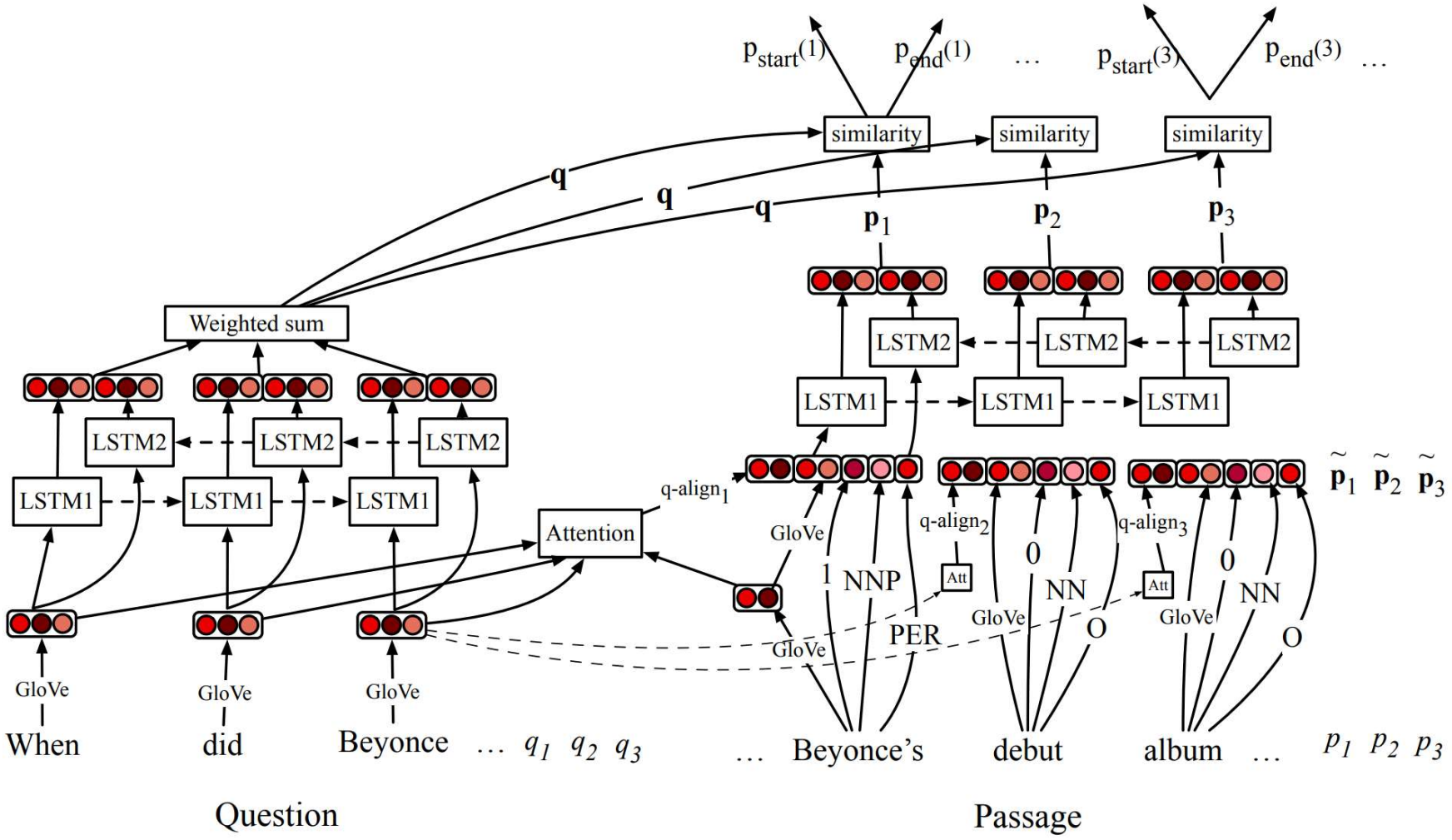


SQuAD 1.1 Results (single model, 2017)

	F1
Logistic regression	51.0
Fine-Grained Gating (Carnegie Mellon U)	73.3
Match-LSTM (Singapore Management U)	73.7
DCN (Salesforce)	75.9
BiDAF (UW & Allen Institute)	77.3
Multi-Perspective Matching (IBM)	78.7
ReasoNet (MSR Redmond)	79.4
DrQA (Chen et al. 2017)	79.4
r-net (MSR Asia) [Wang et al., ACL 2017]	79.7
Human performance	91.2

The Attentive Reader++

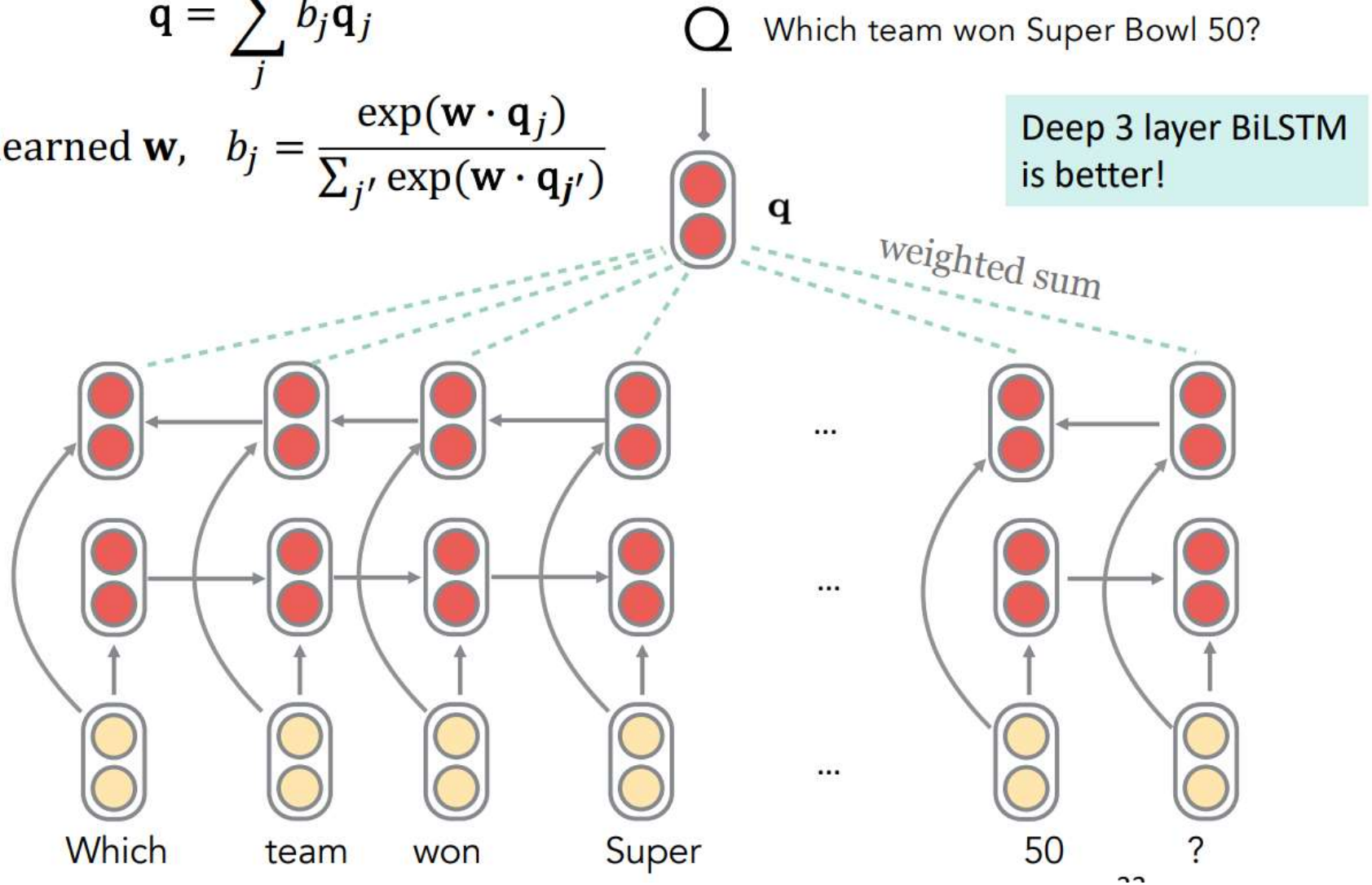
Training Objective: $L = -\sum \log P^{(start)}(a_{start}) - \sum logP^{(end)}(a_{end})$



The Attentive Reader++

$$\mathbf{q} = \sum_j b_j \mathbf{q}_j$$

For learned $\mathbf{w}$, $b_j = \frac{\exp(\mathbf{w} \cdot \mathbf{q}_j)}{\sum_{j'} \exp(\mathbf{w} \cdot \mathbf{q}_{j'})}$



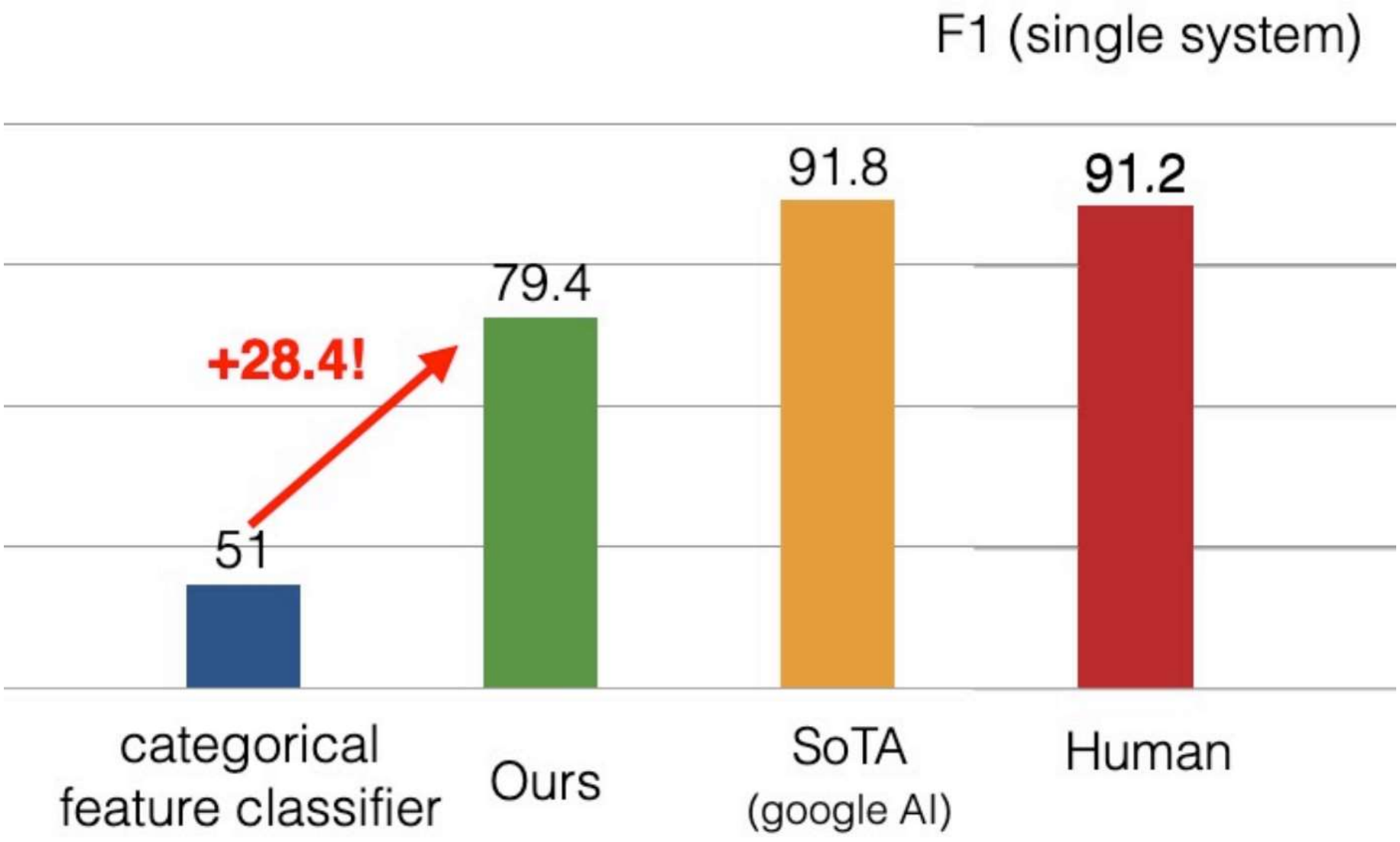
The Attentive Reader++

$\boldsymbol{p}_i$ Vector representation of each token in passage made from concatenation of

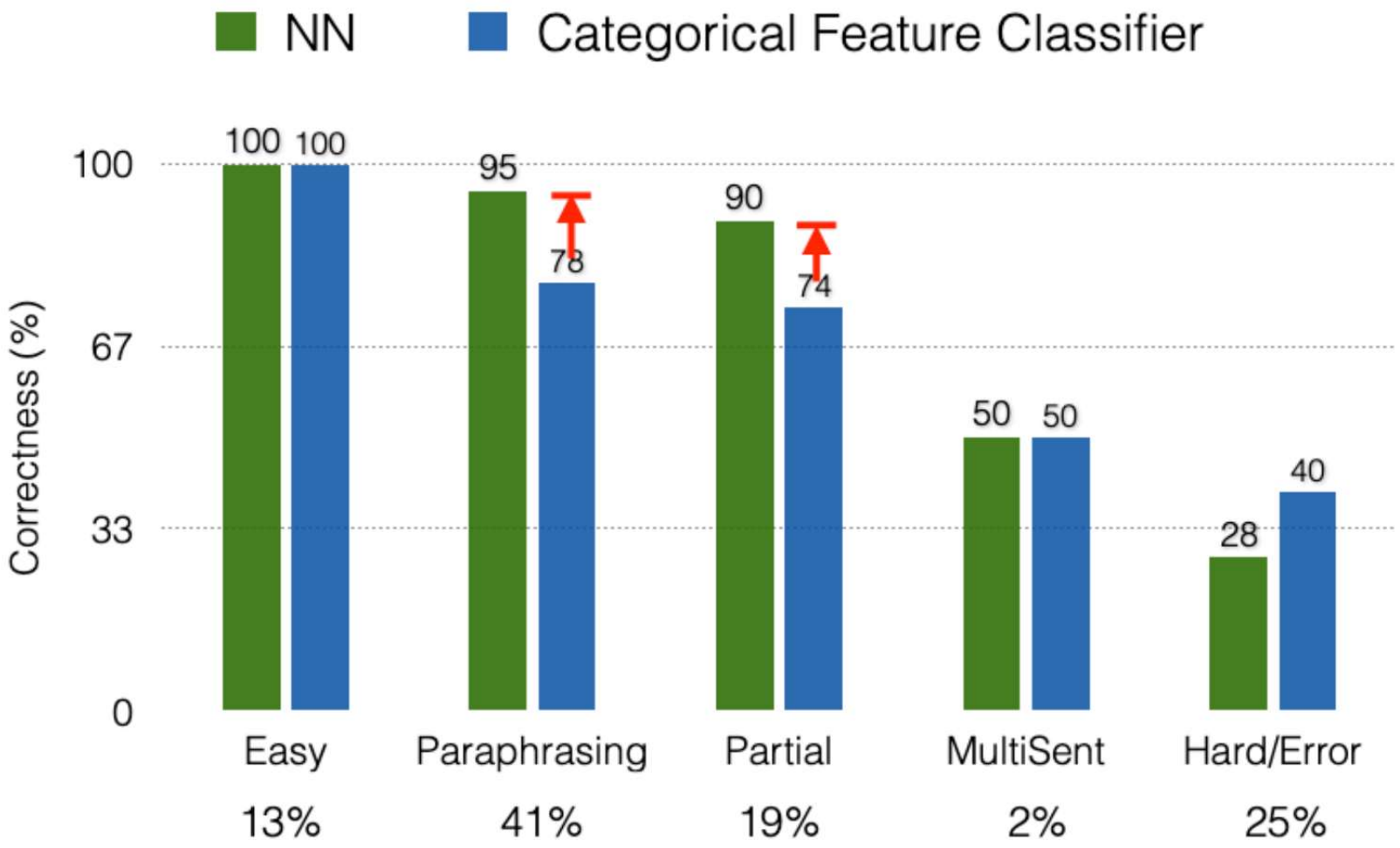
- Word embedding (GloVe 300d)
- Linguistic features: POS & NER tags, one-hot encoded
- Term frequency (unigram probability)
- Exact match: whether the word appears in the question
  - 3 binary features: exact, uncased, lemma
- Aligned question embedding (“car” vs “vehicle”)

$$f_{align}(p_i) = \sum_j a_{i,j}\mathbf{E}(q_j) \qquad q_{i,j} = \frac{\exp(\alpha(\mathbf{E}(p_i)) \cdot \alpha(\mathbf{E}(q_j)))}{\sum_{j'} \exp(\alpha(\mathbf{E}(p_i)) \cdot \alpha(\mathbf{E}(q'_j)))}$$

A big win for neural models

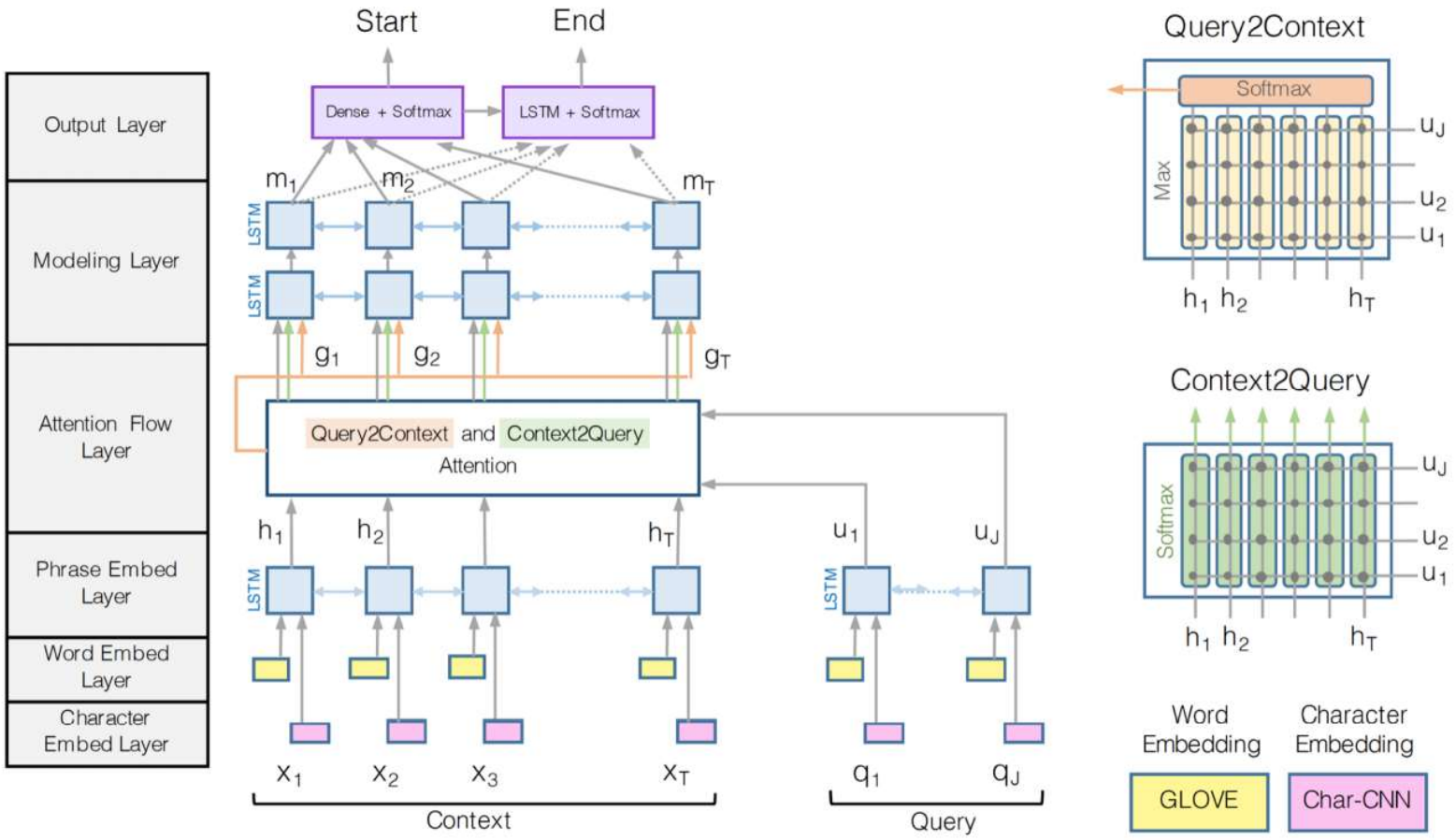


What do these neural mode



BiDAF

- Bi-Directional Attention Flow for Machine Comprehension (Seo, Kembhavi, Farhadi, Hajishirzi, ICLR 2017)



BiDAF

- There are variants of and improvements to the BiDAF architecture over the years, but the central idea is the Attention Flow layer
- Idea: attention should flow both ways – from the context to the question and from the question to the context
- Make similarity matrix (with w of dimension 6d):

$$\boldsymbol{S}_{ij} = \boldsymbol{w}_{\text{sim}}^T[\boldsymbol{c}_i; \boldsymbol{q}_j; \boldsymbol{c}_i \circ \boldsymbol{q}_j] \in \mathbb{R}$$

- Context-to-Question (C2Q) attention: (which query words are most relevant to each context word)

$$\alpha^i = \text{softmax}(\boldsymbol{S}_{i,:}) \in \mathbb{R}^M \quad \forall i \in \{1, \ldots, N\}$$

$$\boldsymbol{a}_i = \sum_{j=1}^{M} \alpha_j^i \boldsymbol{q}_j \in \mathbb{R}^{2h} \quad \forall i \in \{1, \ldots, N\}$$

BiDAF

- Attention Flow Idea: attention should flow both ways – from the context to the question and from the question to the context
- Question-to-Context (Q2C) attention: the weighted sum of the most important words in the context with respect to the query – slight asymmetry through max

$$\boldsymbol{m}_i = \max_j \boldsymbol{S}_{ij} \in \mathbb{R} \quad \forall i \in \{1, \ldots, N\}$$

$$\beta = \text{softmax}(\boldsymbol{m}) \in \mathbb{R}^N$$

$$\boldsymbol{c}' = \sum_{i=1}^{N} \beta_i \boldsymbol{c}_i \in \mathbb{R}^{2h}$$

- For each passage position, output of BiDAF layer is:

$$\boldsymbol{b}_i = [\boldsymbol{c}_i; \boldsymbol{a}_i; \boldsymbol{c}_i \circ \boldsymbol{a}_i; \boldsymbol{c}_i \circ \boldsymbol{c}'] \in \mathbb{R}^{8h} \quad \forall i \in \{1, \ldots, N\}$$

BiDAF

- There is then a “modelling” layer:
  - Another deep (2-layer) BiLSTM over the passage
- And answer span selection is more complex:
  - Start index:

    - Pass output of BiDAF and modelling layer concatenated to a dense FF layer and then a softmax
  - End index:

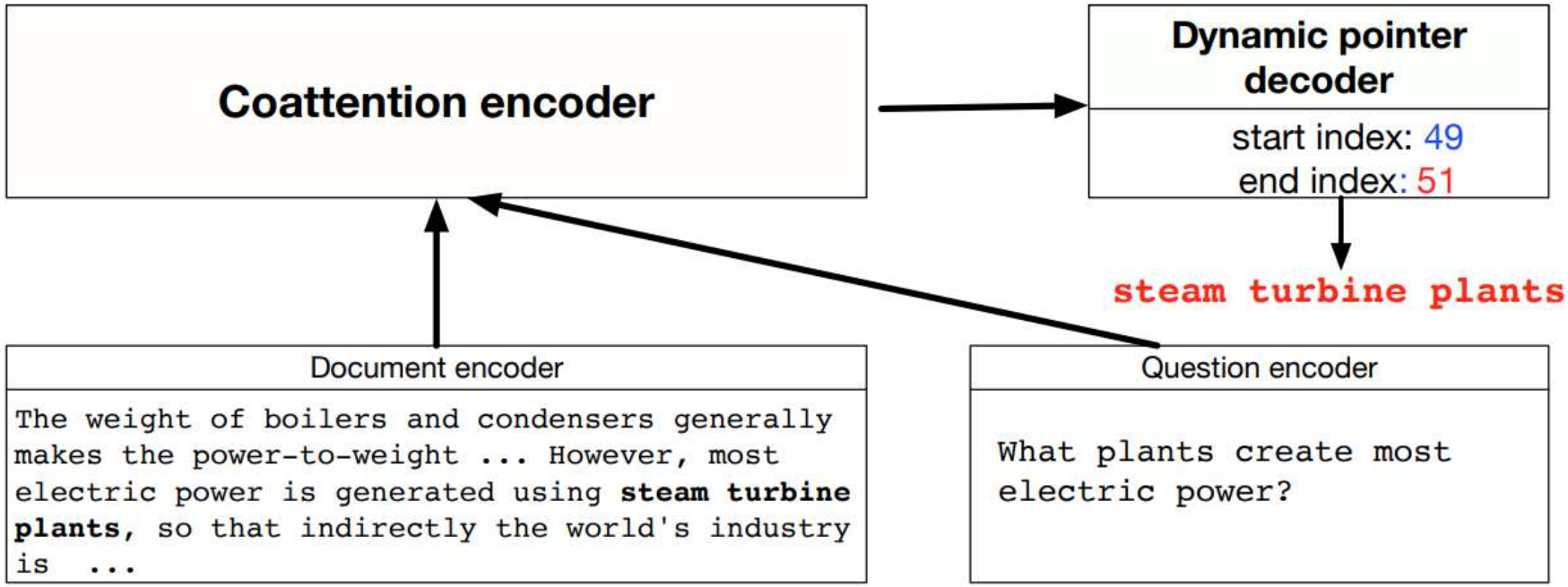
    - Put output of modelling layer M through another BiLSTM to give M2 and then concatenate with BiDAF layer and again put through dense FF layer and a softmax

More advanced architectures

- Most of the work in 2016, 2017, and 2018 employed progressively more complex architectures with a multitude of variants of attention – often yielding good task gains

Dynamic Co-attention Networks for Question Answering

- Flaw: Questions have input-independent representations
- Inter-dependence needed for a comprehensive QA model

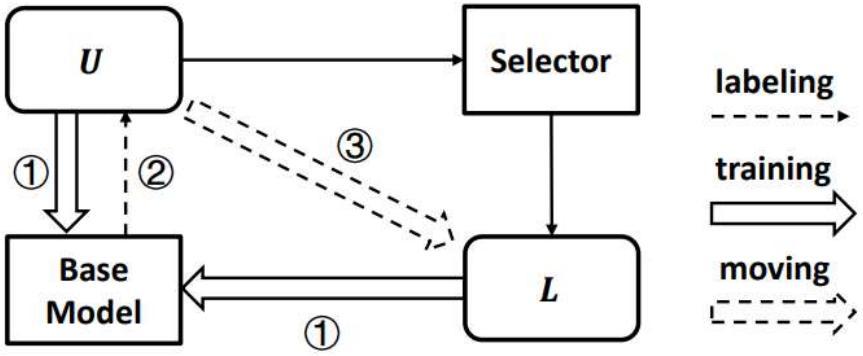


Co-attention: Results on SQUAD Competition

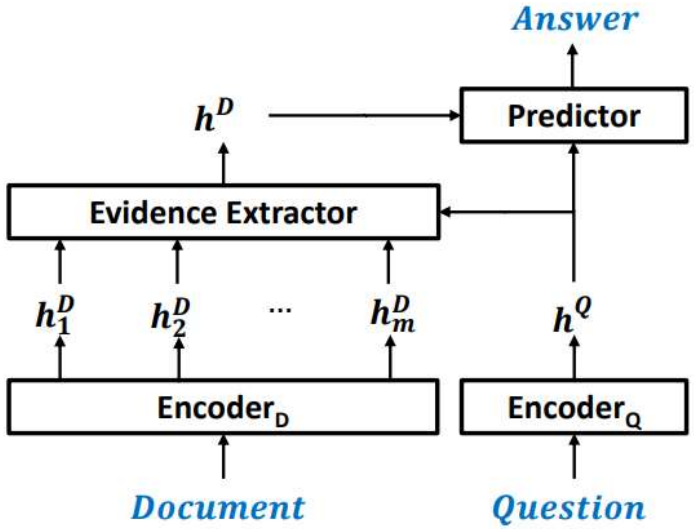
Model	Dev EM	Dev F1
Dynamic Coattention Network (DCN)		
pool size 16 HMN	**65.4**	**75.6**
pool size 8 HMN	64.4	74.9
pool size 4 HMN	65.2	75.2
DCN with 2-layer MLP instead of HMN	63.8	74.4
DCN with single iteration decoder	63.7	74.0
DCN with Wang & Jiang (2016b) attention	63.7	73.7

A self-training method for machine reading comprehension (ACL 2020)

- supervises the evidence extractor with auto-generated evidence labels in an iterative process



Overview of Self-Training MRC



Overall structure of a base model

A self-training method for machine reading comprehension (ACL 2020)

Model / Dataset	RACE-M		RACE-H		MultiRC			DREAM	
	Dev	Test	Dev	Test	Dev			Dev	Test
	Acc	Acc	Acc	Acc	$F1_m$	$F1_a$	EM_0	Acc	Acc
GPT+DPL	64.2	62.4	58.5	60.2	70.5	67.8	13.3	57.3	57.7
BERT-MLP	66.2	65.5	61.6	59.5	71.8	69.1	21.2	63.9	63.2
BERT-HA	67.8	68.2	62.6	60.4	70.1	68.1	19.9	64.2	62.8
BERT-HA+RL	68.5	66.9	62.5	60.0	72.1	69.5	21.1	63.1	63.4
BERT-HA+Rule	66.6	66.4	61.6	59.0	69.5	66.7	17.9	62.5	63.0
BERT-HA+STM	**69.3**‡	**69.2**†	**64.7**‡	**62.6**‡	**74.0**‡	**70.9**‡	**22.0**†	**65.3**‡	**65.8**†
BERT-HA+Gold	N/A	N/A	N/A	N/A	73.7	70.9	27.2	N/A	N/A

Model/Dataset	CoQA	MultiRC					
	P@1	R@1	R@2	R@3	P@1	P@2	P@3
BERT-HA	20.0	28.2	49.8	62.5	62.3	55.2	46.6
+RL	5.2	10.5	22.3	32.9	24.0	25.3	24.7
+Rule	38.4	32.4	53.6	65.1	71.8	59.6	48.7
+STM (iter 1)	32.7	32.8	57.1	70.1	72.2	63.3	52.5
+STM (iter 2)	37.3	**32.9**	**58.0**	**71.3**	**72.7**	**64.4**	**53.5**
+STM (iter 3)	**39.9**	31.4	55.3	68.8	69.5	61.6	51.6
BERT-HA+Gold	53.6	33.7	59.5	73.4	74.5	65.9	54.8

Generating Multi-hop Reasoning Questions to Improve MRC (WWW 2020)

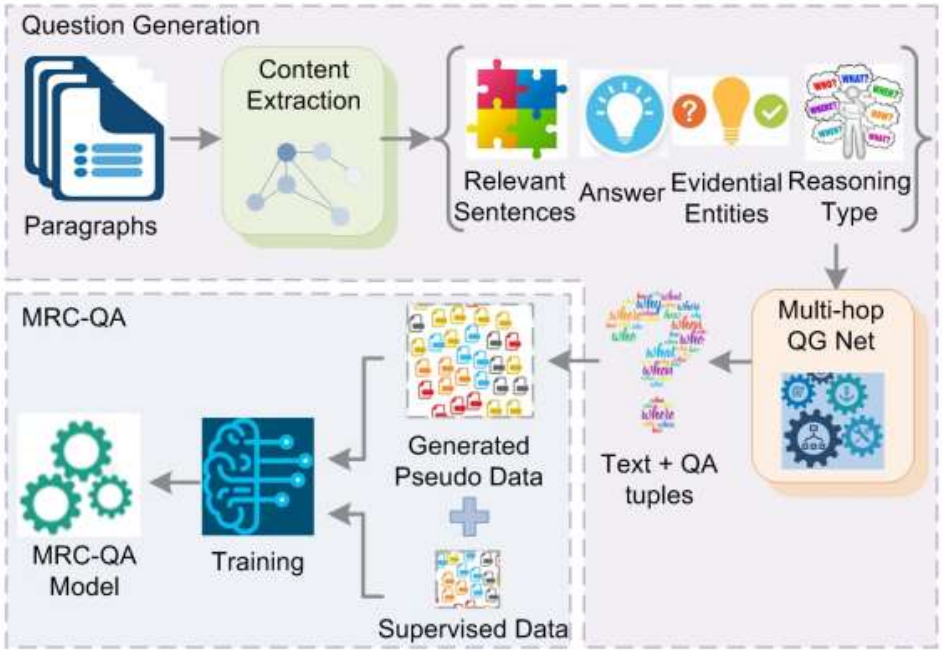
- focuses on the topic of multi-hop question generation
- aims to generate questions needed reasoning over multiple sentences and relations to derive answers

Paragraph 1: (S_1)Shirley Temple Black (April 23, 1928 ~ February 10, 2014) was an American actress, singer, dancer, businesswoman, and diplomat who was Hollywood's number one box-office draw as a child actress from 1935 to 1938. (S_2)As an adult, she was named United States ambassador to Ghana and to Czechoslovakia and also served as Chief of Protocol of the United States.

Paragraph 2: (S_1)Kiss and Tell is a 1945 American comedy film starring then 17-year-old Shirley Temple as Corliss Archer. (S_2)In the film,.... (S_3) The parents' ....

Paragraph 3: ...

Question: What government position was held by the woman who portrayed Corliss Archer in the film Kiss and Tell?
Answer: Chief of Protocol
Supporting sentences: S_1, S_2 in Paragraph 1; S_1 in Paragraph 2.



Generating Multi-hop Reasoning Questions to Improve MRC (WWW 2020)

Methods	All		Bridge		Comparison	
	EM	F1	EM	F1	EM	F1
Task of answer span extraction						
QFE	0.539	0.681	0.541	0.686	0.529	0.661
GRN	0.529	0.667	0.527	0.669	0.539	0.660
Multi-Para	0.456	0.590	0.458	0.595	0.450	0.571
DFGN	0.563	0.697	0.566	0.700	0.552	0.686
Ours	0.613	0.745	0.621	0.753	0.583	0.712
Difference	8.86%	6.90%	9.72%	7.57%	5.62%	3.79%
Task of support sentence prediction						
QFE	0.578	0.845	0.580	0.847	0.567	0.835
GRN	0.524	0.841	0.525	0.839	0.521	0.849
Multi-Para	0.203	0.645	0.206	0.648	0.193	0.631
DFGN	0.515	0.816	0.519	0.819	0.498	0.806
Ours	0.556	0.857	0.564	0.865	0.525	0.827
Difference	7.96%	5.00%	8.67%	5.62%	5.42%	2.61%
Joint task						
QFE	0.346	0.596	0.346	0.599	0.346	0.583
GRN	0.318	0.585	0.321	0.584	0.305	0.589
Multi-Para	0.108	0.402	0.111	0.404	0.096	0.392
DFGN	0.336	0.598	0.341	0.601	0.317	0.587
Ours	0.359	0.634	0.367	0.641	0.327	0.605
Difference	6.78%	5.98%	7.62%	6.66%	3.15%	3.07%

Performance on various question types

Support Sentences
S1:Margaret Seeger (born June 17, 1935) is an American folksinger. **S2**:She is also well known in Britain, where she has lived for more than 30 years, and was married to the singer and songwriter Ewan MacColl until his death in 1989. **S3**:James Henry Miller (25 January 1915 ∼ 22 October 1989), better known by his stage name Ewan MacColl, was an English folk singer, songwriter, communist, labour activist, actor, poet, playwright and record producer.
Gold Standard
Answer: American **Question**: What nationality was James Henry Miller's wife?
Ours Method
Answer: Britain **Question**: Where has the wife of James Henry Miller lived for more than 30 years? **Answer**: English **Question**: What nationality was the husband of Margaret Seeger? **Answer**: folksinger **Question**: What is the occupation for the wife of James Henry Miller? **Answer**: American **Question**: What nationality was the wife of James Henry Miller?

Case studies on QG model