Scaling up Reading Comprehension

Eunsol Choi

November 2017

Question Answering from Raw Text

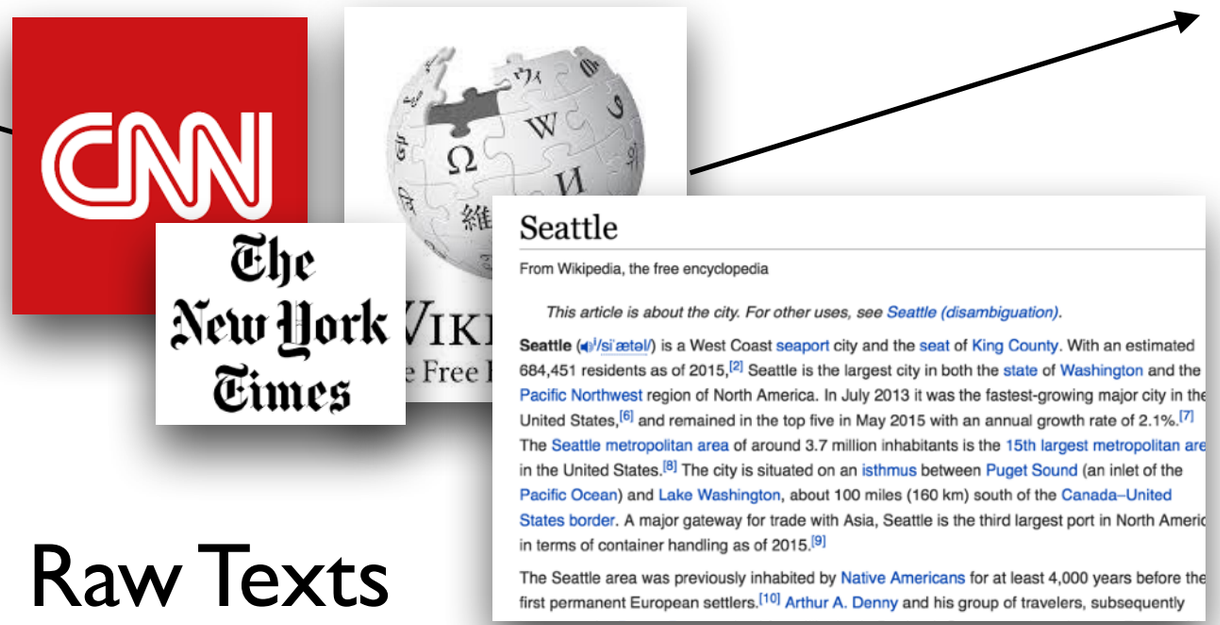
Reading Comprehension

Related Dataset:

WikiQA (Yang et al 15)
CNN dataset (Hermann et al 14)
Children Book Test (Hill et al 15)
SQuAD (Rajpurkar et al 16)
TriviaQA (Joshi et al 17)

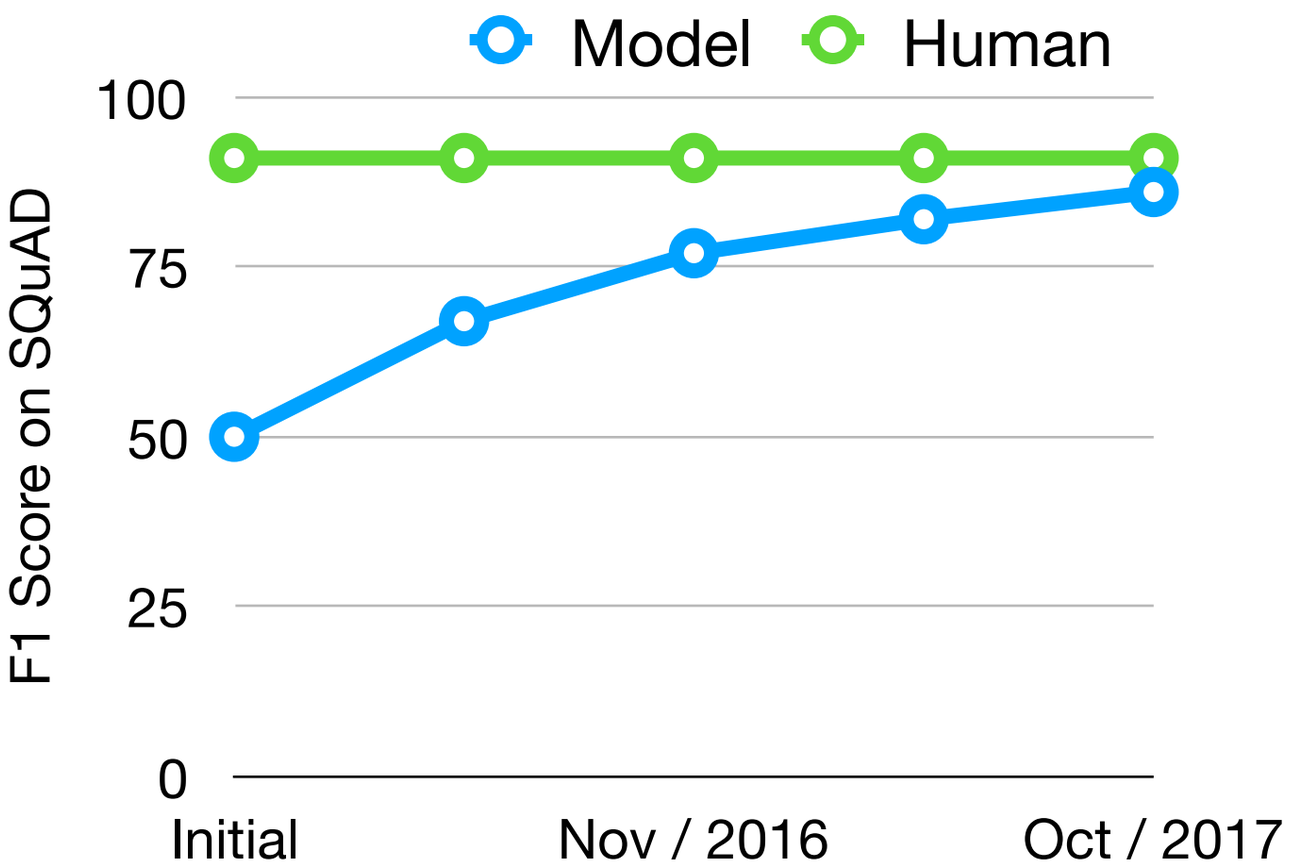
Query

Answer

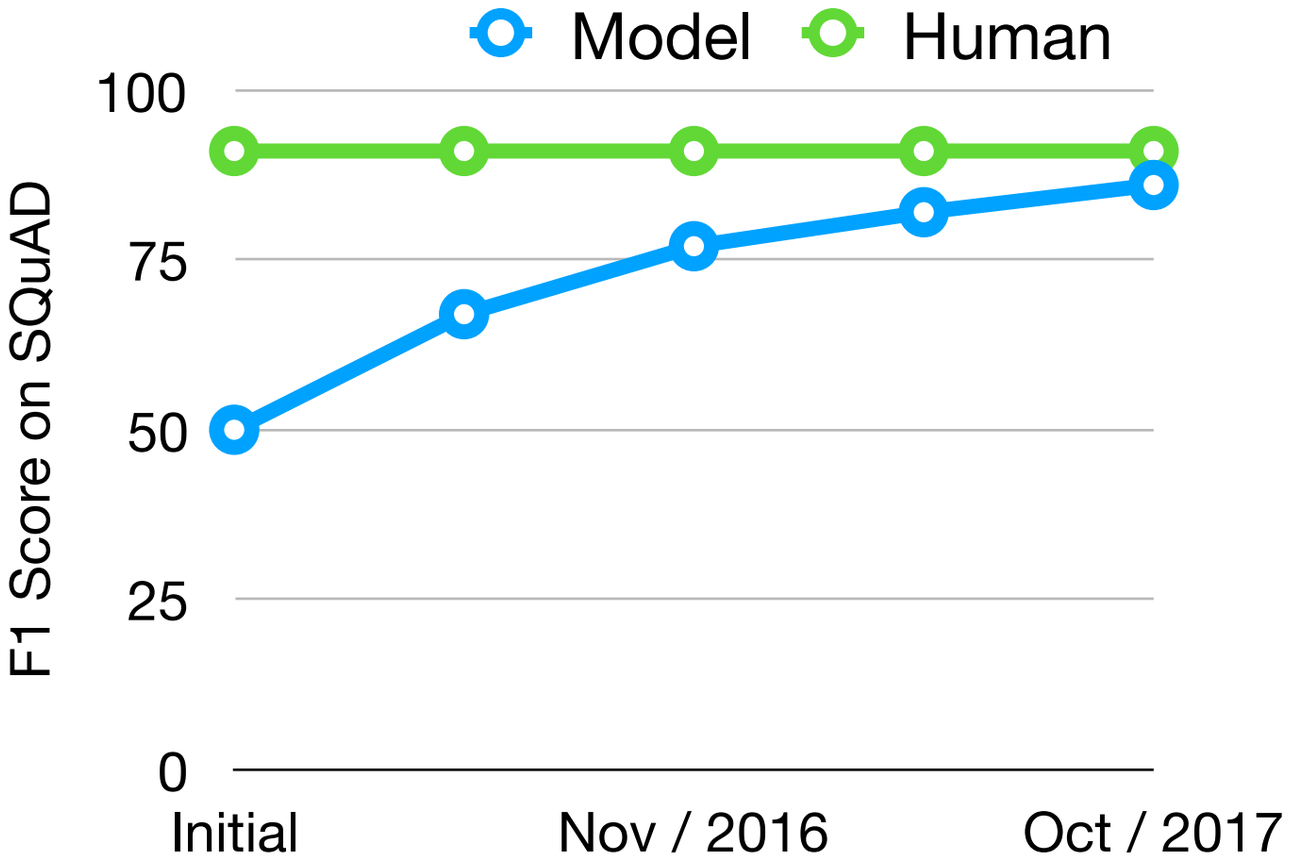


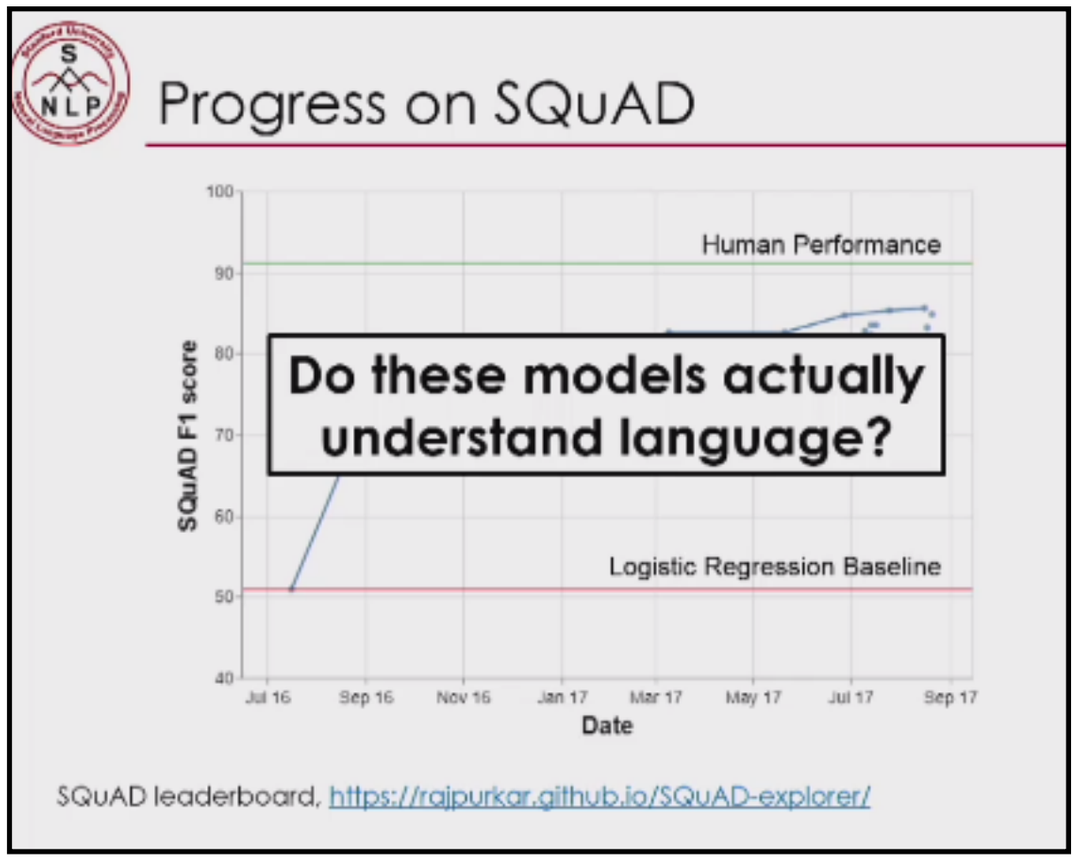
Raw Texts

Recent Progress in Reading Comprehension



Recent Progress in Reading Comprehension

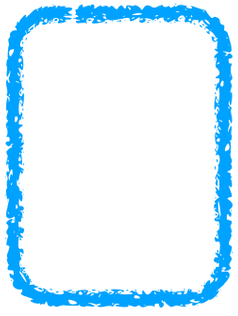




Jia and Liang EMNLP17

Did we solve reading comprehension already?

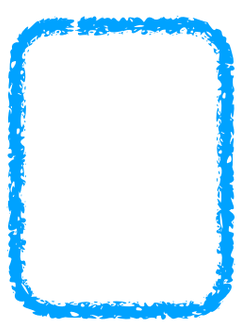
Improving
Model



Coarse-to-Fine Question Answering For Long Document

[Choi et al, ACL 17]

Improving
Model



Coarse-to-Fine Question Answering For Long Document [Choi et al, ACL 17]

Improving
Data

TriviaQA: A Challenge Dataset for Reading Comprehension [Joshi et al, ACL 17]

Improving Model

Coarse-to-Fine Question Answering For Long Document [Choi et al, ACL 17]

Improving Data

TriviaQA: A Challenge Dataset for Reading Comprehension [Joshi et al, ACL 17]

Applying Model

Reading Comprehension for Relation Extraction [Levy et al, CoNLL17]

Introduction

Improving Model

Coarse-to-Fine Question Answering For Long Document [Choi et al, ACL 17]

Improving Data

TriviaQA: A Challenge Dataset for Reading Comprehension [Joshi et al, ACL 17]

Applying Model

Reading Comprehension for Relation Extraction [Levy et al, CoNLL17]

Future Work

Introduction

Improving Model

Coarse-to-Fine Question Answering For Long Document [Choi et al, ACL 17]

Improving Data

TriviaQA: A Challenge Dataset for Reading Comprehension [Joshi et al, ACL 17]

Applying Model

Reading Comprehension for Relation Extraction [Levy et al, CoNLL17]

Future Work

Coarse-to-Fine Question Answering For Long Documents

Eunsol Choi, Daniel Hewlett, Jakob Uszkoreit
Illia Polosukhin, Alexandre Lacoste, Jonathan Berant
ACL 2017

Research Question

- Given a question and a **long document**, how can we efficiently find an answer?

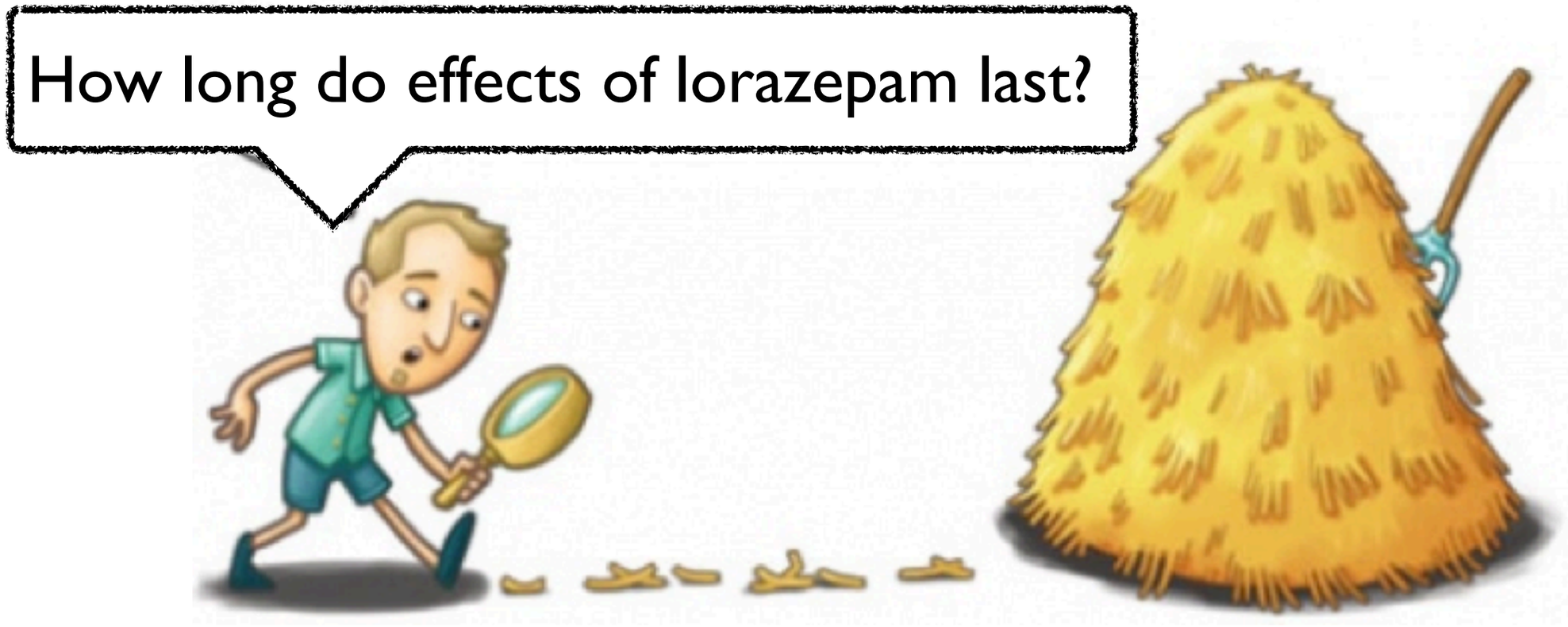
- State-of-the-art recurrent neural network is inappropriate to handle long document:

- Speed: Sequential processing

- Effectiveness: Often forgets earlier sentences

Question Answering

Question Answering



Question Answering

How long do effects of lorazepam last?

Lorazepam

From Wikipedia, the free encyclopedia

Not to be confused with Loprazolam.

Lorazepam, sold under the brand name **Ativan** among others, is a benzodiazepine medication often used to treat anxiety disorders.[4] Lorazepam reduces anxiety, interferes with new memory formation, reduces agitation, induces sleep, treats seizures, treats nausea and vomiting, and relaxes muscles.[5][6] Lorazepam is used for the short-term treatment of anxiety, trouble sleeping, acute seizures including status epilepticus, sedation of people in hospital, as well as sedation of aggressive patients.[6][7][8][9] Due to tolerance and dependence, lorazepam is recommended for short-term use, up to two to four weeks only.

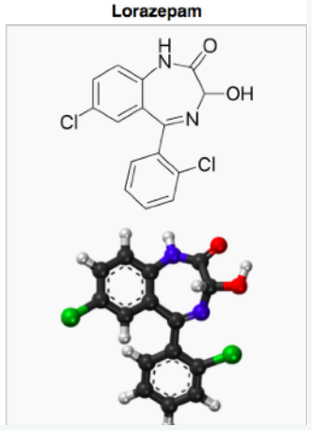
Among benzodiazepines, lorazepam has possible physical addiction potential.[5] Lorazepam also has misuse potential; the main types of misuse are for recreational purposes.[10] Long-term effects of benzodiazepines include tolerance, dependence, benzodiazepine withdrawal syndrome, and cognitive impairments which may not completely reverse after stopping treatment. Withdrawal symptoms can range from anxiety and insomnia to seizures and psychosis. Adverse effects, including inability to form new memories, depression, and paradoxical effects, such as excitement or worsening of seizures, may occur. Children and the elderly are more sensitive to the adverse effects of benzodiazepines.[5][11][12] Lorazepam impairs body balance and standing steadiness and is associated with falls and hip fractures in the elderly.[13]

Lorazepam was initially patented in 1963 and went on sale in the United States in 1977.[14] It is on the World Health Organization's List of Essential Medicines, the most important medications needed in a basic health system.[15]

Contents [hide]

1 Medical uses

Lorazepam

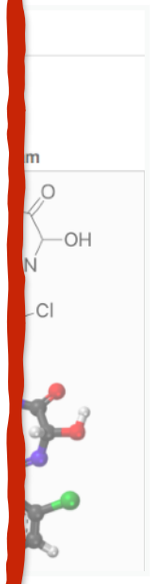


Question Answering

H

Reading document closely from start to the end is probably NOT the best strategy.

Can we search for relevant sentences and read them more carefully?



Task

How long do effects of lorazepam last?

Lorazepam

From Wikipedia, the free encyclopedia

Not to be confused with Loprazolam.

Lorazepam, sold under the brand name **Ativan** among others, is a benzodiazepin
Lorazepam reduces anxiety, interferes with new memory formation, reduces agita
and vomiting, and relaxes muscles.[5][6] Lorazepam is used for the short-term trea
including status epilepticus, sedation of people in hospital, as well as sedation of a
dependence, lorazepam is recommended for short-term use, up to two to four wee

Among benzodiazepines, lorazepam has possible physical addiction potential.[5] L
of misuse are for recreational purposes.[10] Long-term effects of benzodiazepines
withdrawal syndrome, and cognitive impairments which may not completely revers
can range from anxiety and insomnia to seizures and psychosis. Adverse effects,
depression, and paradoxical effects, such as excitement or worsening of seizures,
sensitive to the adverse effects of benzodiazepines.[5][11][12] Lorazepam impairs b
associated with falls and hip fractures in the elderly.[13]

Lorazepam was initially patented in 1963 and went on sale in the United States in
List of Essential Medicines, the most important medications needed in a basic hea

Contents [hide]

1 Medical uses

Contents [hide]

Task

How long do effects of lorazepam last?

Lorazepam

From Wikipedia, the free encyclopedia

Not to be confused with Loprazolam.

Lorazepam, sold under the brand name **Ativan** among others, is a benzodiazepin
Lorazepam reduces anxiety, interferes with new memory formation, reduces agita
and vomiting, and relaxes muscles.[5][6] Lorazepam is used for the short-term treat
including status epilepticus, sedation of people in hospital, as well as sedation of a
dependence, lorazepam is recommended for short-term use, up to two to four wee

Among benzodiazepines, lorazepam has possible physical addiction potential.[5] L
of misuse are for recreational purposes.[10] Long-term effects of benzodiazepines
withdrawal syndrome, and cognitive impairments which may not completely revers
can range from anxiety and insomnia to seizures and psychosis. Adverse effects,
depression, and paradoxical effects, such as excitement or worsening of seizures,
sensitive to the adverse effects of benzodiazepines.[5][11][12] Lorazepam impairs b
associated with falls and hip fractures in the elderly.[13]

Lorazepam was initially patented in 1963 and went on sale in the United States in
List of Essential Medicines, the most important medications needed in a basic hea

Contents [hide]

1 Medical uses

Contents [hide]

1 Medical uses
2 Adverse effects
- 2.1 Contraindications
- 2.2 Special groups and situations
- 2.3 Tolerance and dependence
- 2.4 Withdrawal
- 2.5 Interactions
- 2.6 Overdose
- 2.7 Detection in body fluids

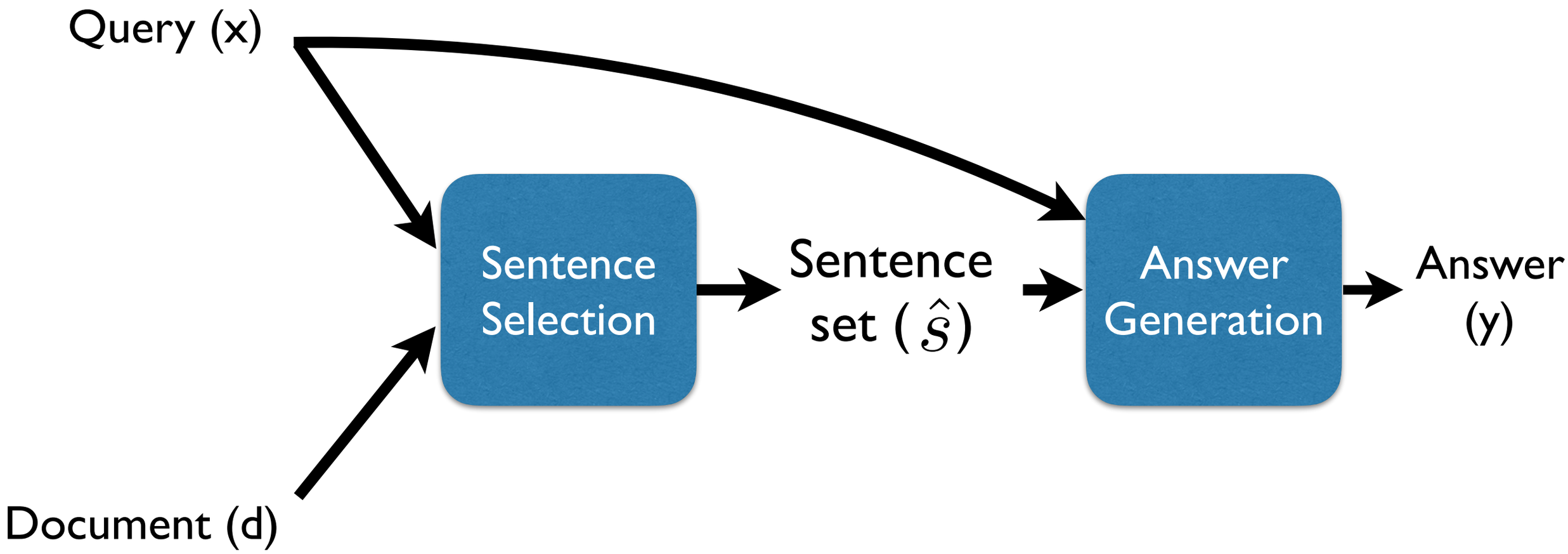
3 Pharmacology
- 3.1 Pharmacokinetics
- 3.2 Pharmacodynamics

4 History

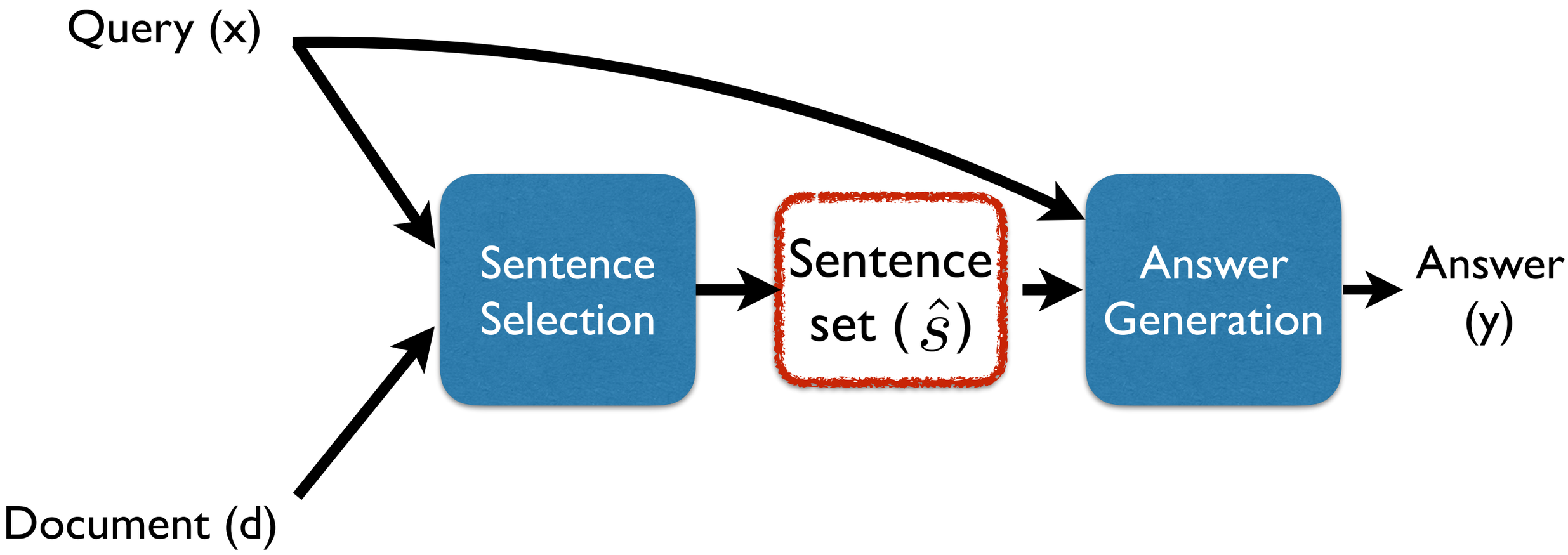
receiving a continuous lorazepam infusion.[88] Intravenous injections should be given slowly and patients closely monitored for side effects, such as respiratory depression, hypotension, or loss of airway control.

Peak effects roughly coincide with peak serum levels,[79] which occur 10 minutes after intravenous injection, up to 60 minutes after intramuscular injection, and 90 to 120 minutes after oral administration,[73][79] but initial effects will be noted before this. A clinically relevant lorazepam dose will normally be effective for six to 12 hours, making it unsuitable for regular once-daily administration, so it is usually prescribed as two to four daily doses when taken regularly, but this may be extended to five or six, especially in the case of elderly patients

Coarse-to-Fine Approach



Coarse-to-Fine Approach



This Work

Query (x)

- Coarse-to-Fine model for question answering
- Substantially faster (up to 6.7 times) with comparable accuracies
- Learning without direct supervision for evidence sentence

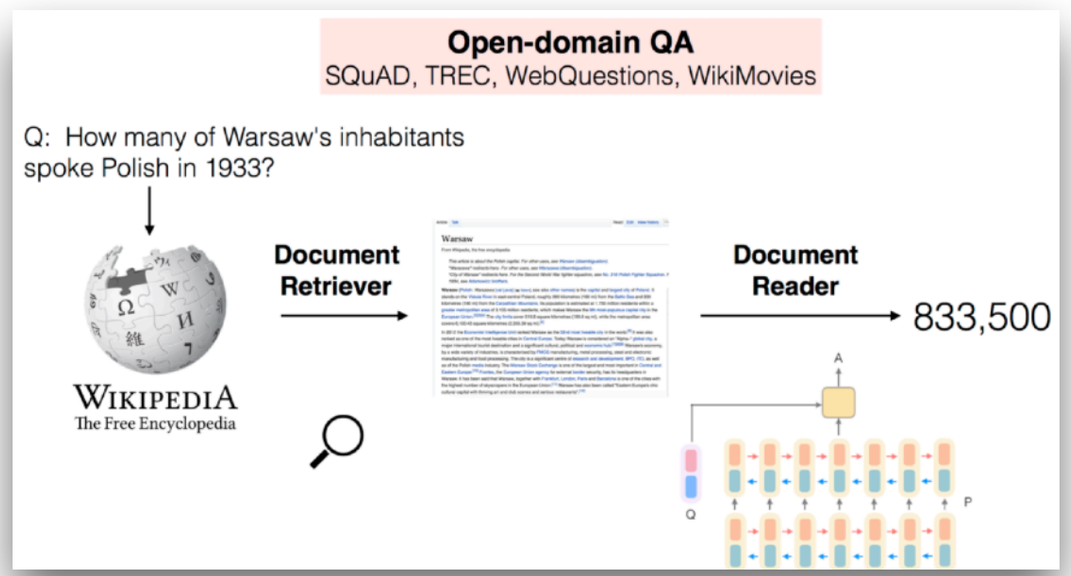
nswer (y)

Documer

Related Work

- Coarse-to-Fine model for different applications
  (Charniak et al, 06, Cheng and Lapata, 16, Yang et al, 16, Lei et al, 16)

- Two-staged processes for question answering:
  (Servelyn and Mochitti 15, Yang et al, 16, Jurczyk et al, 16, dos Santos et al, 16, Sultan et al 16, Chen et al, 17)



Data

- WikiReading (Hewlett et al, ACL 16)
  - Wikipedia InfoBox

- WikiReading-Long (Hewlett et al, ACL 16)
  - Challenging WikiReading subset, longer documents

- WikiSuggest (Choi et al, ACL 17)
  - Query suggest from Google, answered by Google snippets

WikiReading

- Taken from Wikipedia.
- Infobox properties and article.

Entity	Property	Document	Answer
Folkart Towers	Country	Folkart Towers are twin skyscrapers in Turkish city of Izmir.	Turkey
Canada	Located next to body of water	Canada is a country… extended from the Atlantic to the Pacific and northward into the Artic Ocean	Atlantic Ocean, Pacific, Arctic Ocean
Breaking Bad	Start time	Breaking Bad is a TV series… from January 20, 2008	20 January 2008

WikiReading

- Taken from Wikipedia.
- Infobox properties and article.

✓ Large-scale dataset covering various domains

✓ Most documents are relatively short (480 words)

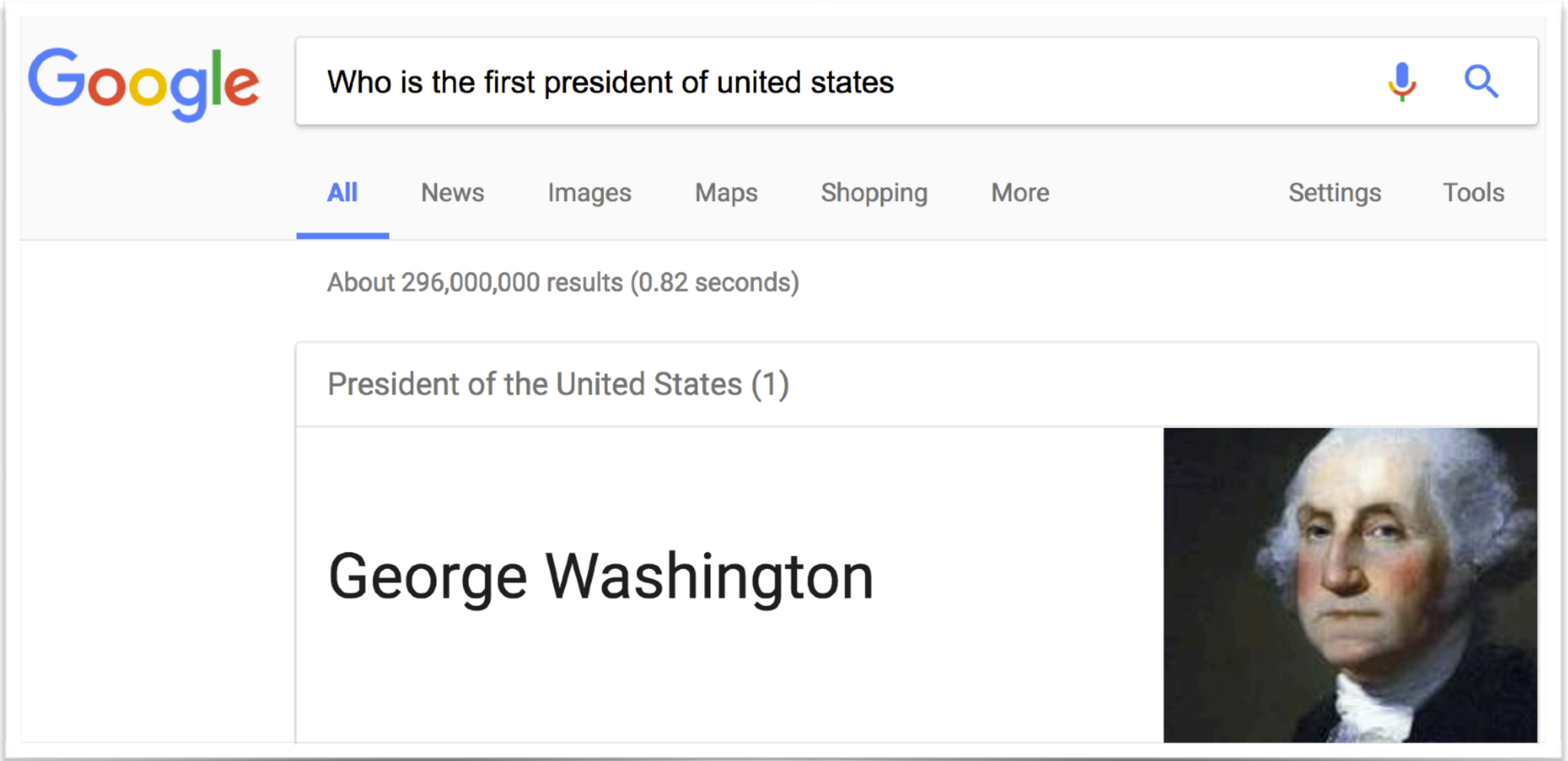
Entity			
Folka			
Canada	to body of water	the Atlantic to the Pacific and northward into the Artic Ocean	ean, Pacific, Arctic Ocean
Breaking Bad	Start time	Breaking Bad is a TV series… from January 20, 2008	20 January 2008

WikiReading-Long

- Pruned to have documents with at least 10 sentences
- Contains 1.97 million instances (~ 15% of original data)
- Single document contains 1200 words on average

WikiSuggest

- Question from Google's user queries
- Answers from Google's auto suggested answer
- Document from Wikipedia



WikiSuggest Examples

Query	Answer
how many officials in a nfl football game	seven officials
the 11th tarot card	Major Arcana
what age ronald reagan become president	69 years
ohio basketball coach	Saul Phillips
how old is ed marinaro	born on March 31, 1950
allers syndrome	Ehlers-Danlos

WikiSuggest Examples

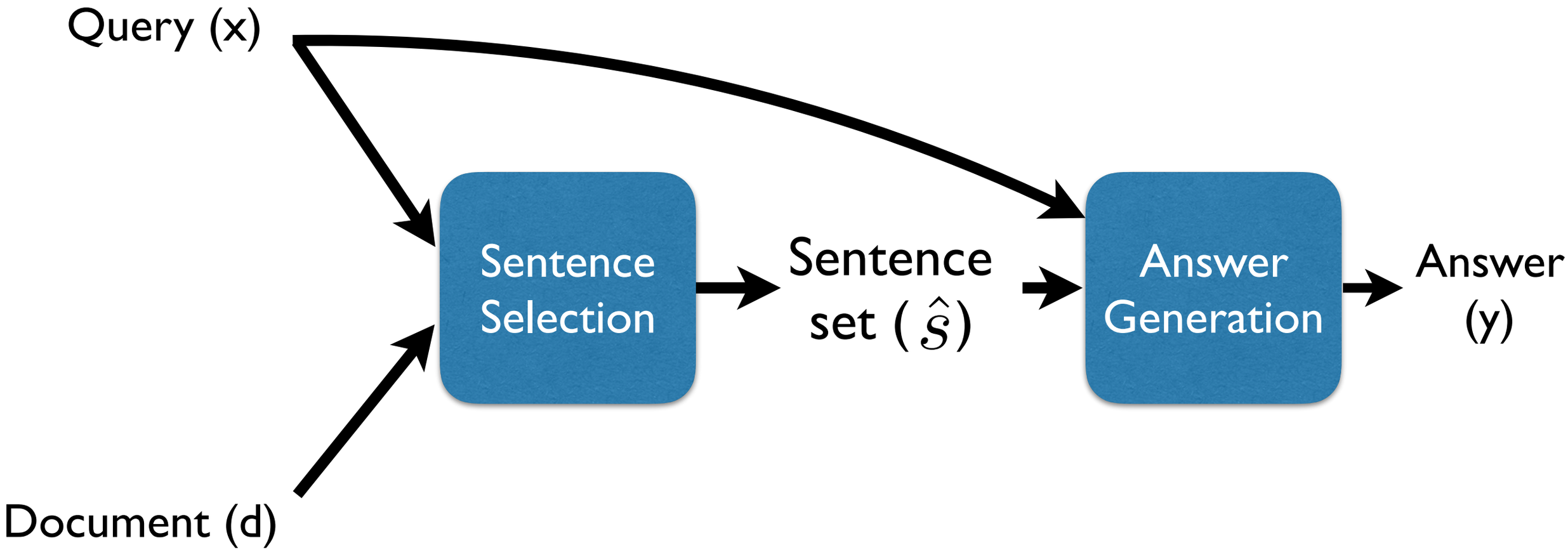
Qu	
hov gan	
the	
wha pre	
ohi	
how [illegible]	born on March 31, 1950
allers syndrome	Ehlers-Danlos

- ✓ Large-scale, noisy dataset covering various domain (3.5M)
- ✓ More diverse and natural questions
- ✓ Including systematically generated noise (~25%)

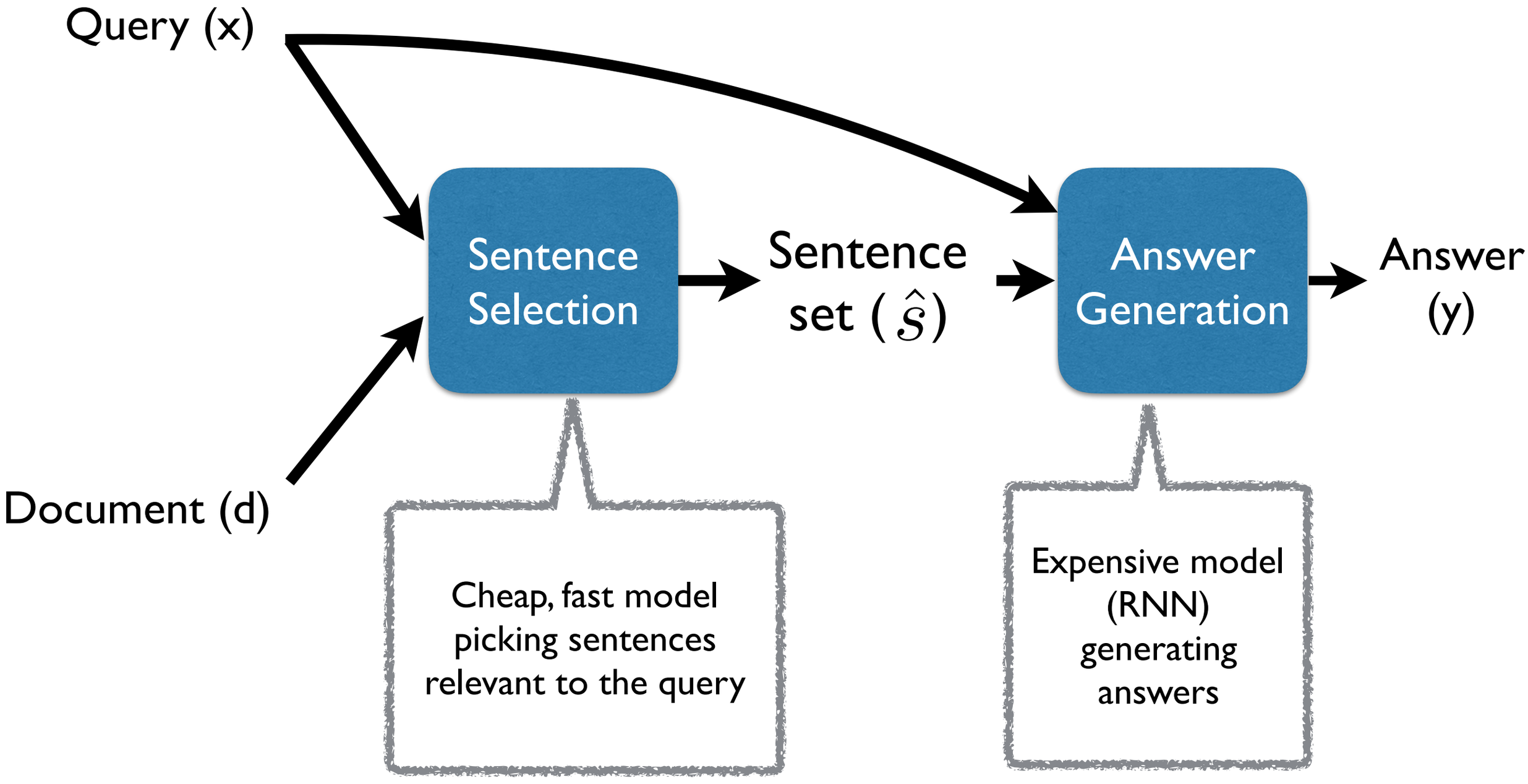
Dataset Summary

	# Examples	# Unique Queries	# of tokens / doc
WikiReading	18M	867	0.5K
WikiReadingLong	2M	239	1.2K
WikiSuggest	3.5M	3.5M	5.9K

Model


Query (x)
Sentence Selection
Sentence set ($\hat{S}$)
Answer Generation
Answer (y)
Document (d)

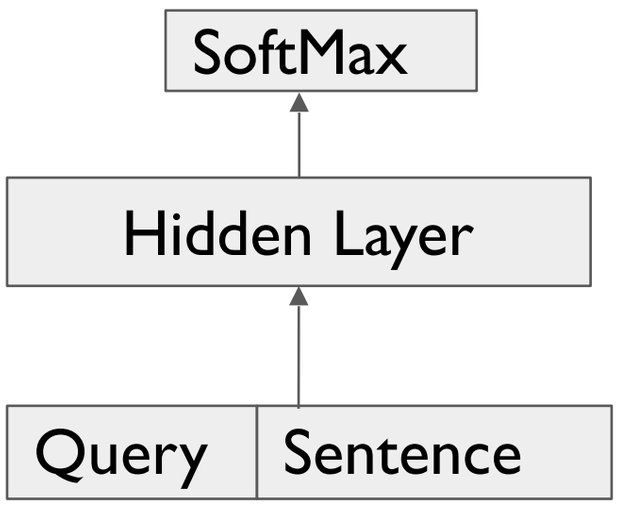
Model


Query (x)
Document (d)
Sentence Selection
Sentence set ($\hat{s}$)
Answer Generation
Answer (y)
Cheap, fast model picking sentences relevant to the query
Expensive model (RNN) generating answers

Sentence Selection Model

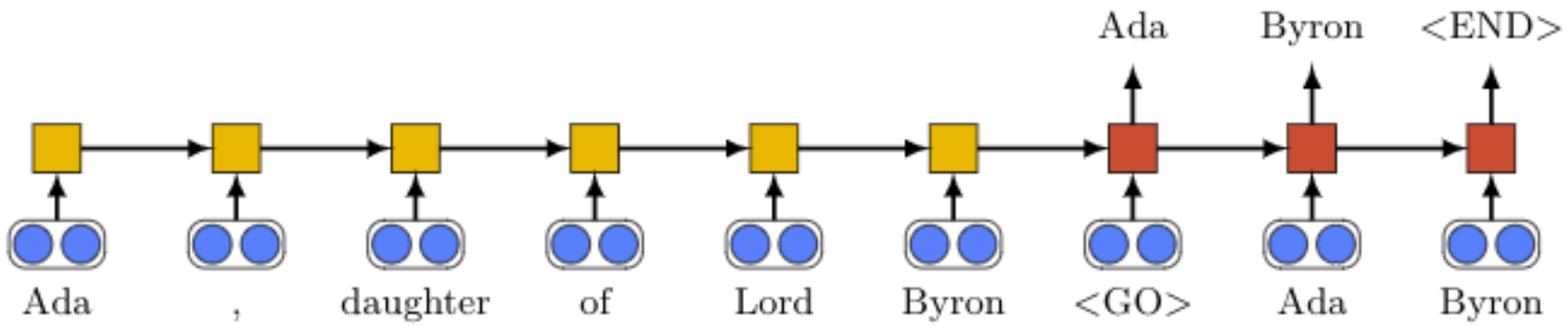
- Take query and document as input
- Coarse and fast sentence representation (BoW)
- Computes relevance score for each sentence ($P(s|x,d)$)

  to generate sentence set to pass on to answer generation model.

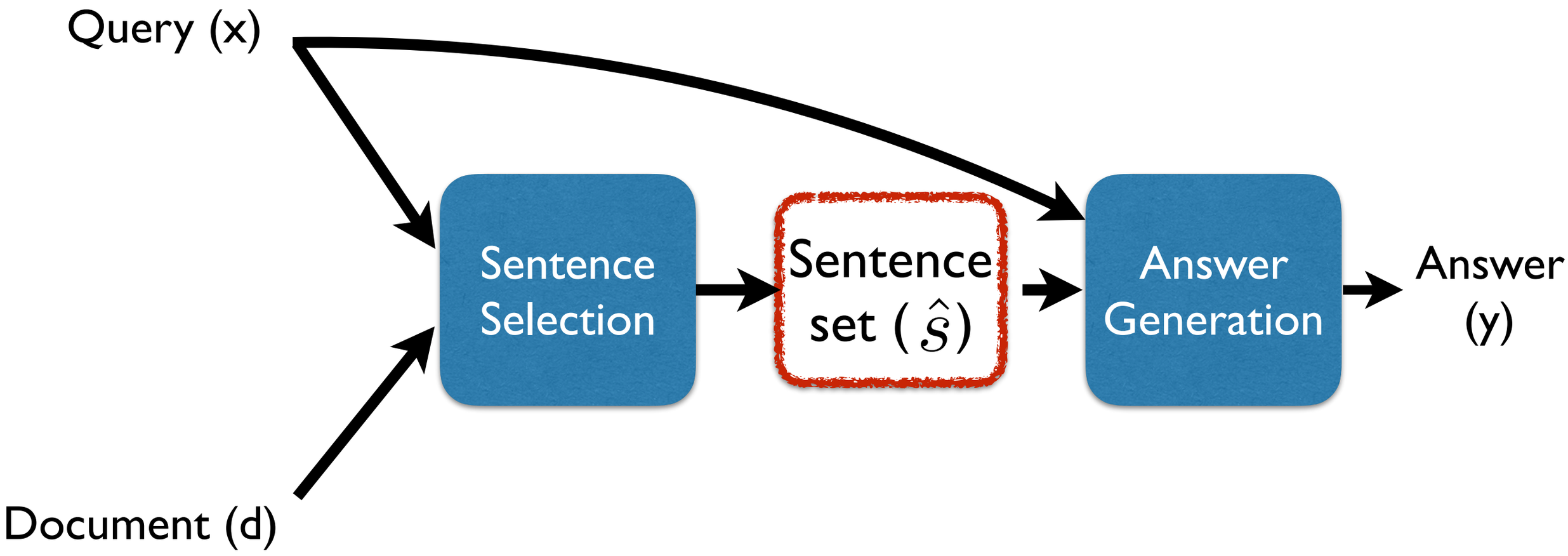


Answer Generation Model
(Hewlett et al, ACL16)

- Given a sentence set and a query, generate answer strings
- RNN encoder-decoder model with placeholders



Learning



We do not have supervision for which sentences contain the information.

Can we know which sentence contains answer?

- Good heuristics:
  - Sentence with an answer string is the sentence that you should pay close attention to.

Q: Folkart Towers, country	A: Turkey	S: Folkart Towers are twin skyscrapers in Turkish city of Izmir.
Q: Where did Alexandro Friedmann die?	A: St. Petersburg	S: Alexandro Friedmann was born in St. Petersburg.

Answer String Match Statistics

	Answer String Exists	Avg. # of Answer Match	Answer in First Sentence if answer exists
WikiReading	**47.1%**	1.22	75.1%
WikiReading Long	**50.4%**	2.18	31.3%
WikiSuggest	100.0%	**13.95**	33.6%

False Negative

False Positive

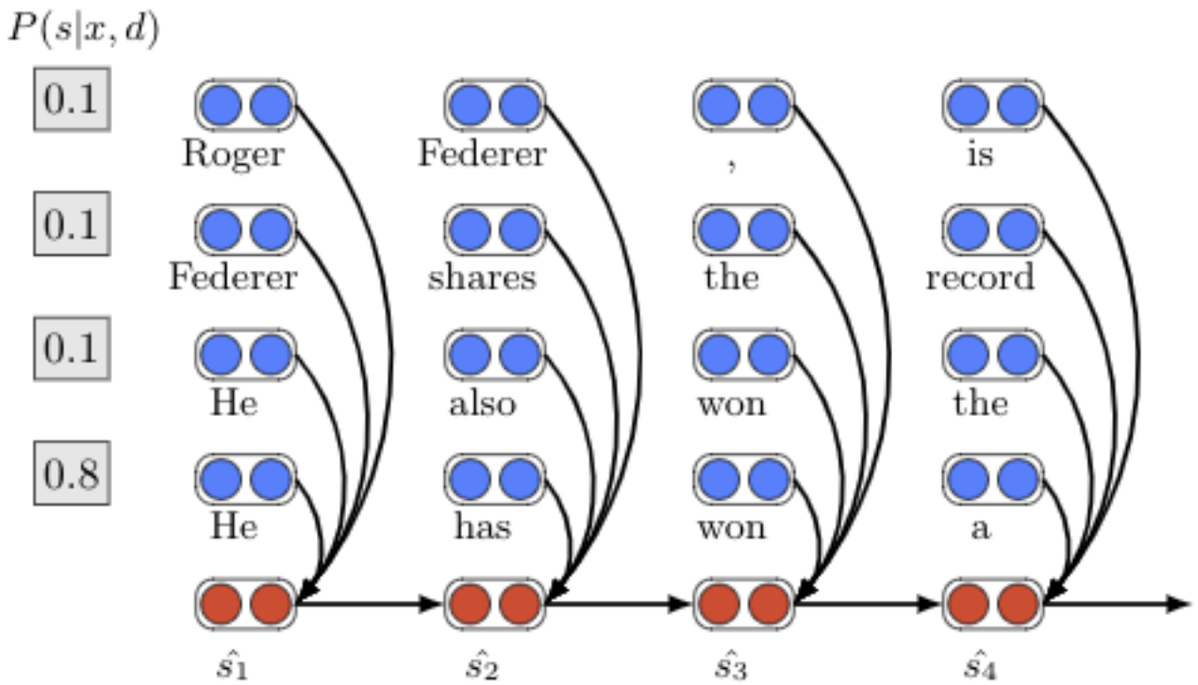
Pipeline (Distant Supervision)

- Separate objective for two models

$$\log p(s^*|x, d) + \log p(y^*|x, s^*)$$

- Gold sentence (s^*): First sentence with an answer string match or first sentence if answer string match does not exist.

Soft Attention



- Make a "blended" token representation by merging each sentence token weighted by its relevancy score p(s|x).
- Allows end-to-end learning.

$$\log p(y^*|x,d) = \log p(y^*|x,\hat{s})$$

Hard Attention
(Reinforcement Learning)

- Action: Choosing a sentence

- Reward: Log probability of answer with chosen sentence

$$R(s) = \log P(y^*|s, x)$$

$$E[R] = \sum_s P(s|x) \cdot R(s)$$

$$= \sum_s P(s|x) \cdot \log P(y^*|s, x)$$

- Can approximate the gradient with sampling (REINFORCE)

$$\nabla \log P(y^*|\tilde{s}, x) + \log P(y^*|\tilde{s}, x) \cdot \nabla \log P(\tilde{s}|x)$$

Hard Attention
(Reinforcement Learning)

- Can be flexible on the number of sentences to pass on to the answer generation model

- Curriculum learning (Ross et al, AISTAT11)

- Trained with pipeline objective at the beginning

Evaluation

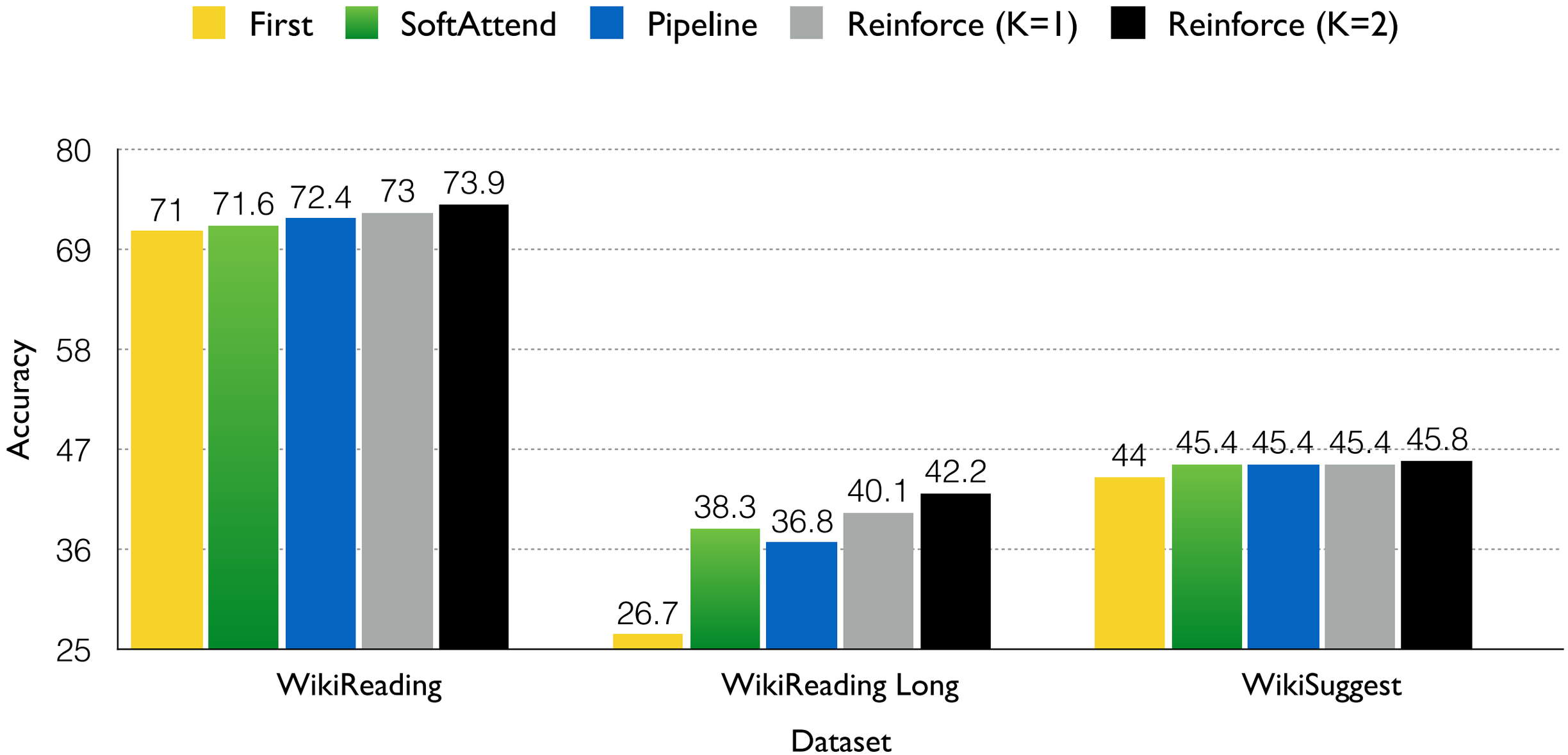
- Answer accuracy:
  - exact match accuracy

- Efficiency:
  - time to finish document encoding

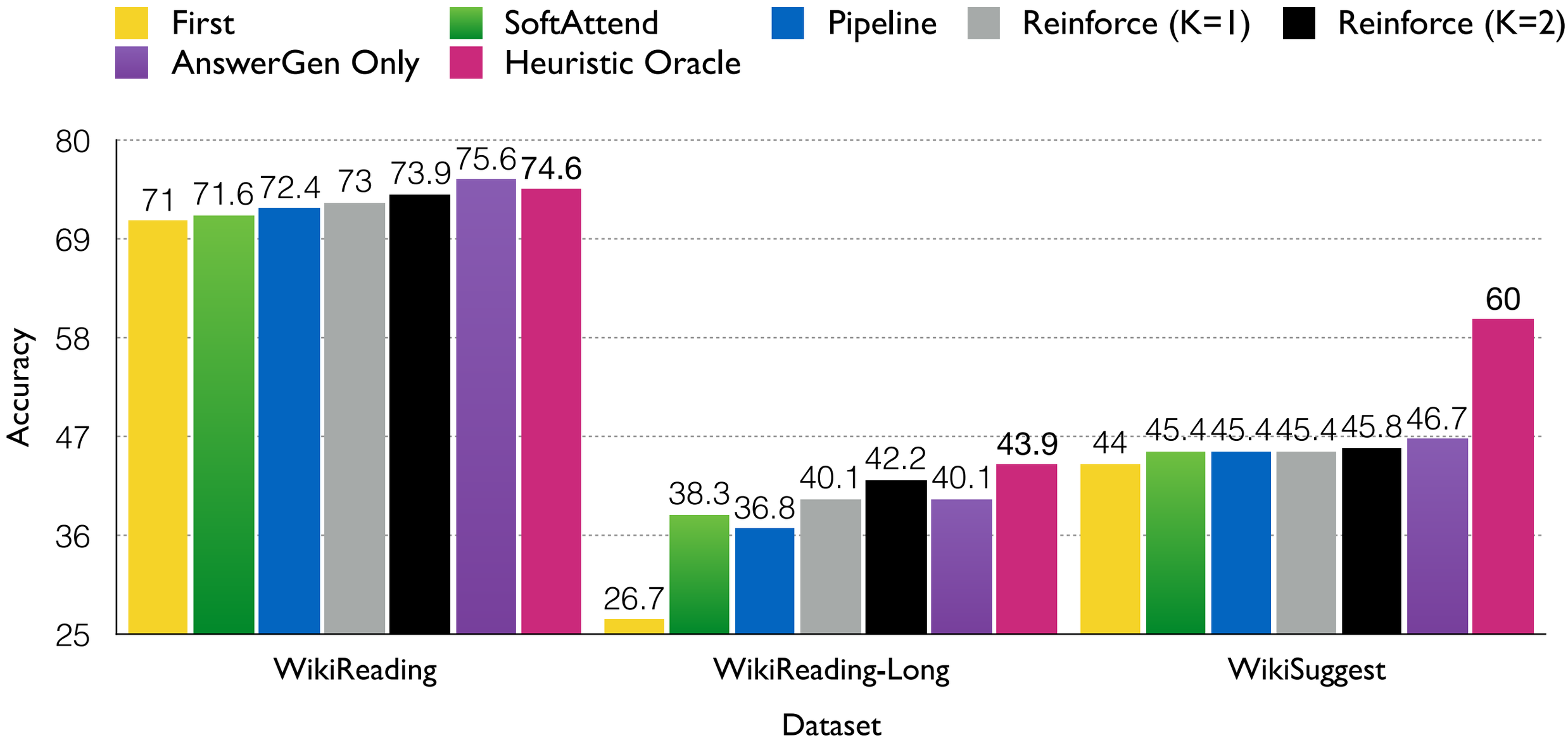
Comparison Systems

- First sentence baseline
- Answer generation baseline:
  - Input is the first 300 tokens.
- Heuristic oracle:
  - Input is the sentence with answer string or the first sentence when there is no answer match.

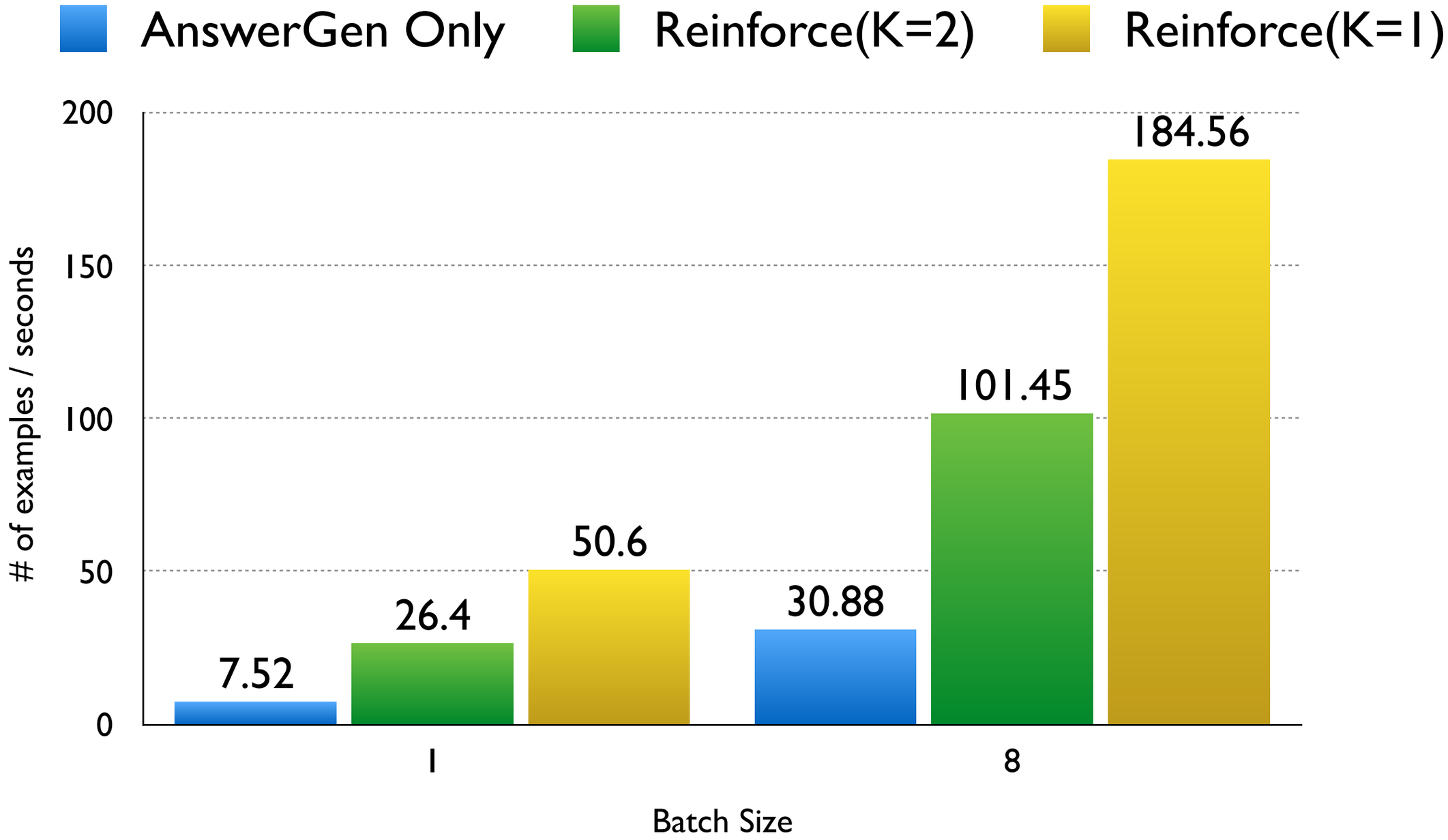
Accuracy Results



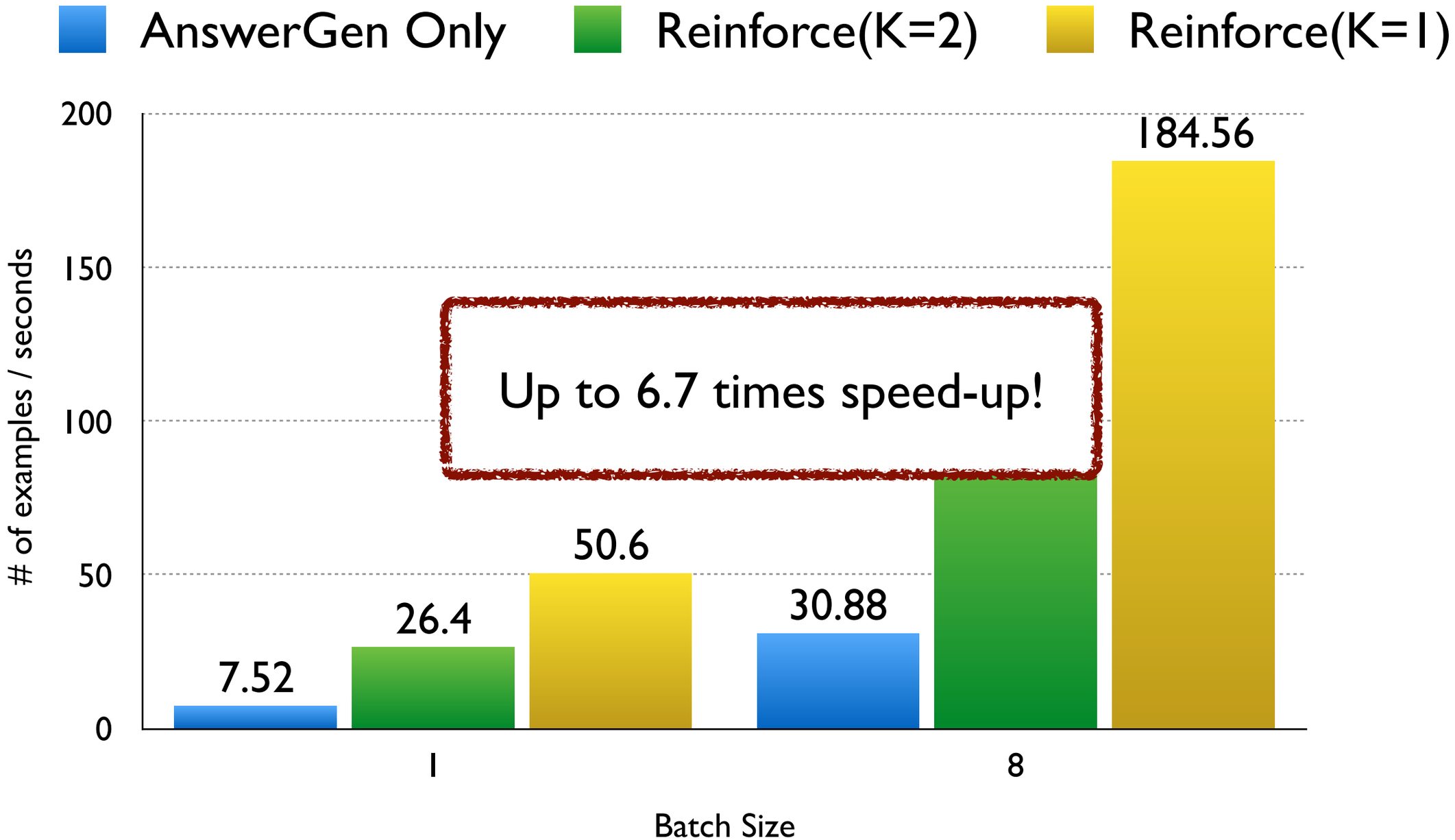
Accuracy Results



Speed Results



Speed Results



Error Analysis

	WikiReading-Long	WikiSuggest
No evidence in document	58%	16%
Error in answer generation	26%	30%
Error in sentence selection	16%	6%
Noisy QA pairs	0%	48%

Conclusion

- Coare-to-Fine model for question answering
- Efficient model (up to 6.7 times speed up) with comparable accuracies
- Learning strategy without direct supervision for evidence sentence

Introduction

Improving Model

Coarse-to-Fine Question Answering For Long Document [Choi et al, ACL 17]

Improving Data

TriviaQA: Challenge Dataset for Reading Comprehension [Joshi et al, ACL 17]

Applying Model

Reading Comprehension for Relation Extraction [Levy et al, CoNLL17]

Future Work

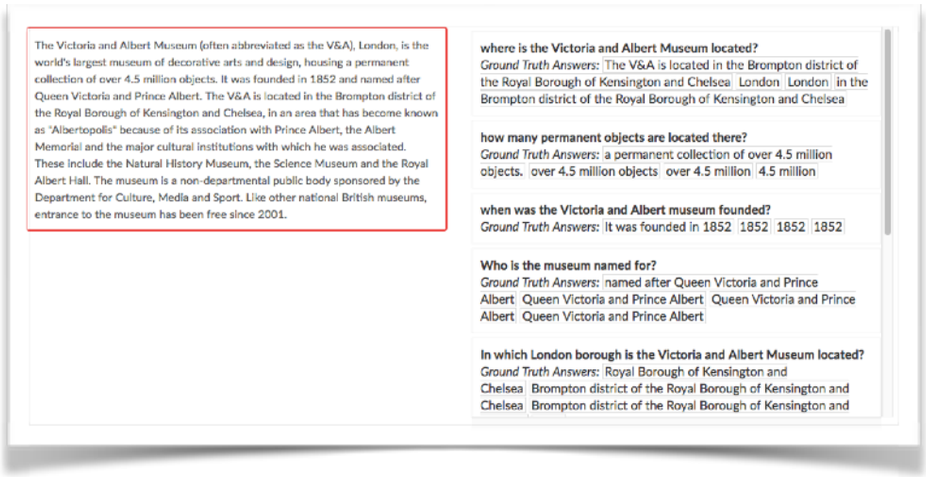
TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension

Mandar Joshi, Eunsol Choi,
Dan Weld, Luke Zettlemoyer
ACL 2017

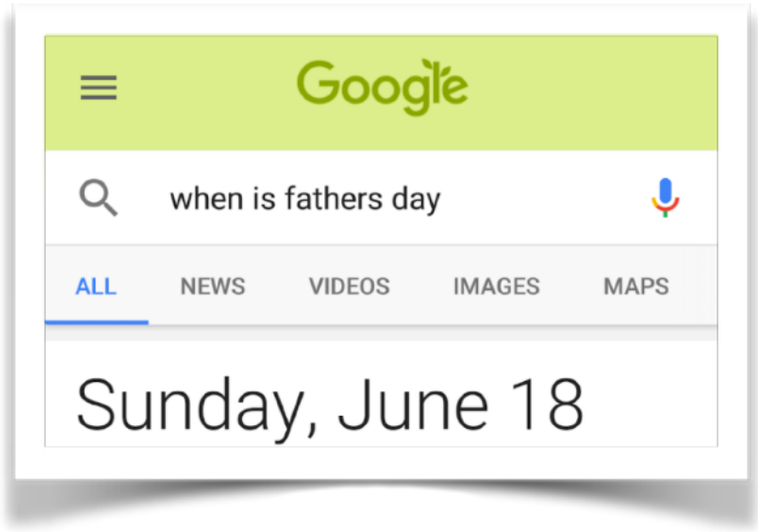
UWNLP

slides by Mandar Joshi

Existing QA Datasets



[Rajpurkar et al 16]



[TREC QA dataset]

paragraph-level

corpus-level

TriviaQA

TriviaQA

- 95K question answer pairs from trivia websites
- 650K documents collected independently
- Average length of document is 2,895 words
- Significant difference between human performance and baselines

Why is TriviaQA challenging?

- Questions require aggregating information from different parts of the document:
  - e.g.) Who was born first, Kiefer Sutherland or Christian Slater?

Why is TriviaQA challenging?

- Questions require aggregating information from different parts of the document:
  - e.g.) Who was born first, Kiefer Sutherland or Christian Slater?
- Questions are often compositional and detailed:
  - e.g.) What was the surname of the woman who was the inspiration behind the Rolling Stones song Angie?

Why is TriviaQA challenging?

- Questions require aggregating information from different P
  - e.g.) What was the surname of the woman who was the inspiration behind the Rolling Stones song Angie?
- Q

Questions involving reasoning across multiple sentences: 40%
Questions involving time frame: 34 %
Average question length: 14 tokens

Dataset Comparisons

Dataset	Dataset Size	Well formed Questions	Freeform Answer
TREC	✗	✓	✓
WikiQA	✗	✗	✗
SQuAD	✓	✓	✓
NewsQA	✓	✓	✓
MS Marco	✓	✗	✓
SearchQA	✓	✓	✓
TriviaQA	✓	✓	✓

Dataset Comparisons

Dataset	Dataset Size	Well formed Questions	Freeform Answer	Varied Domain	Independent Evidence	Lengthy Document
TREC	✗	✓	✓	✓	✓	✓
WikiQA	✗	✗	✗	✓	✓	✓
SQuAD	✓	✓	✓	✗	✗	✗
NewsQA	✓	✓	✓	✗	✗	✓
MS Marco	✓	✗	✓	✓	✓	✗
SearchQA	✓	✓	✓	✓	✓	✗
TriviaQA	✓	✓	✓	✓	✓	✓

Phase 1: Question and answer collection

SFQL - QUIZZES

Tuesday 6th September 2016
Set by Ormskirk

Round 1
1a Spencer Compton in 1743 was the first British Prime Minister to die in o
WILMINGTON
1b Astana is the capital of which Asian country?
KAZAKHSTAN
2a Which novel by Michael Ondaatje shared the Booker Prize in 1992 and
THE ENGLISH PATIENT
2b Who painted Ballet Rehearsal in 1873?
EDGAR DEGAS

PAR COGITIO ET VERBIS

ORRELL & DISTRICT QUIZ LEAGUE

For matches to be played on 4th April 2017
Set by Phatmarkie

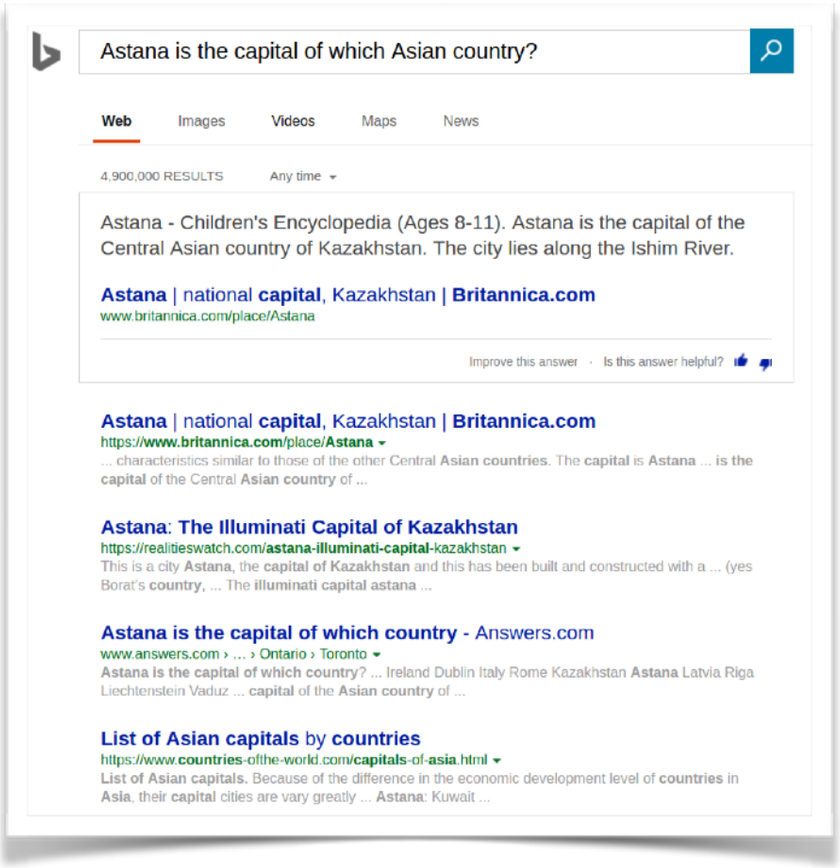
ROUND ONE

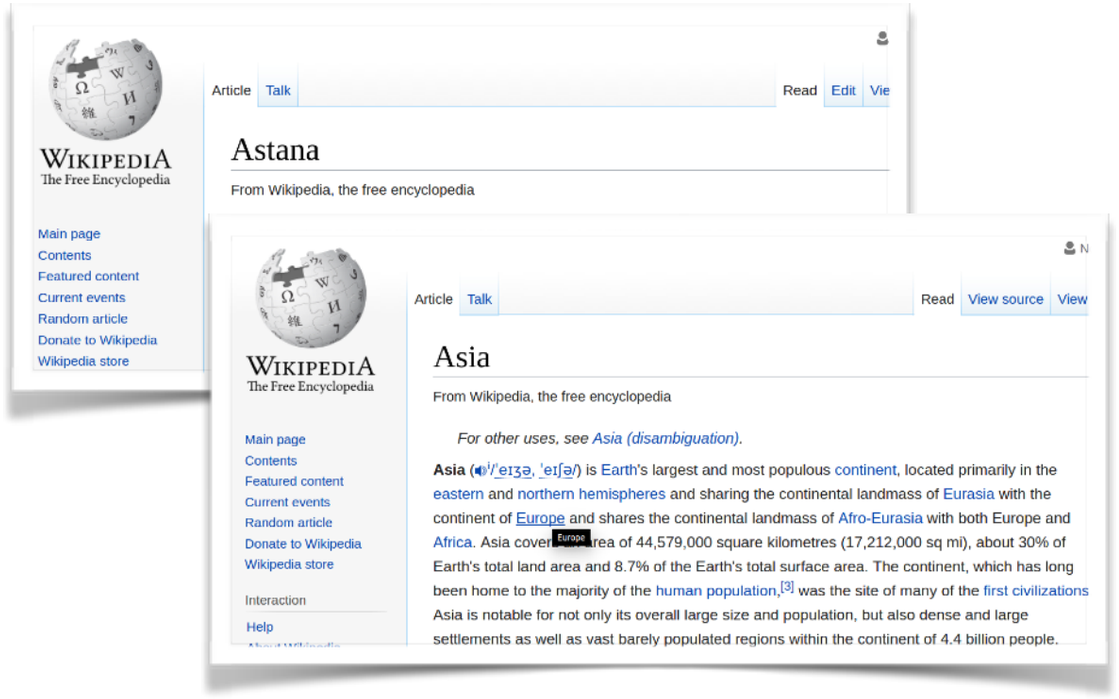
1	Which English queen was born in 1508 at Wulfhall in Wiltshire?	**Jane Seymour**
2	Jay Garrick, Barry Allen and Wally West have all assumed the mantle of which comic book superhero?	**The Flash**
3	What word is used to describe a shoe with a canvas upper body and a sole made of rope?	**Espadrille**
4	Which U.K. road runs for 75 miles from Prescot, Merseyside to Wetherby, North Yorkshire?	**A58**

1. Astana is the capital of which Asian country? KAZAKHSTAN
2. Who painted Ballet Rehearsal in 1873? EDGAR DEGAS
3. ...

Phase 2: Evidence Document Collection

Astana is the capital of which Asian country?





Web: via Search Engine

Wikipedia : via Entity linking

Distant supervision

Astana is the capital of which Asian country?

✓

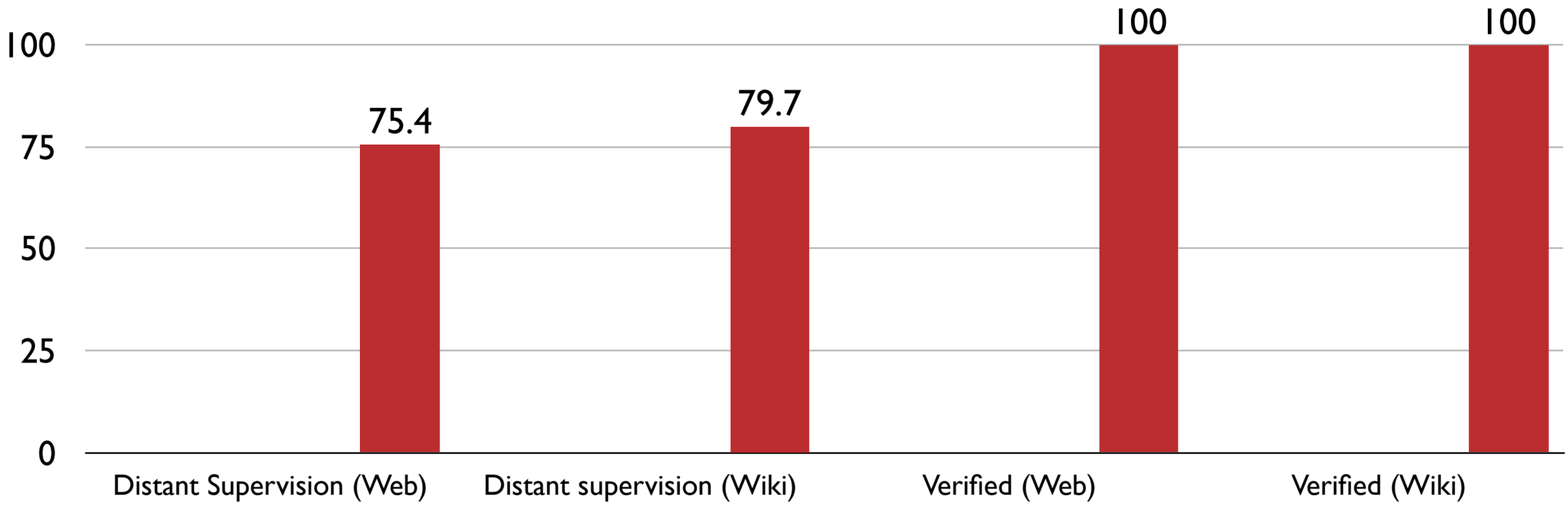
Astana is the capital city of Kazakhstan.

✗

After the dissolution of the Soviet Union and the consequent independence of Kazakhstan, the city's original form was restored in the modified form *Akmola*

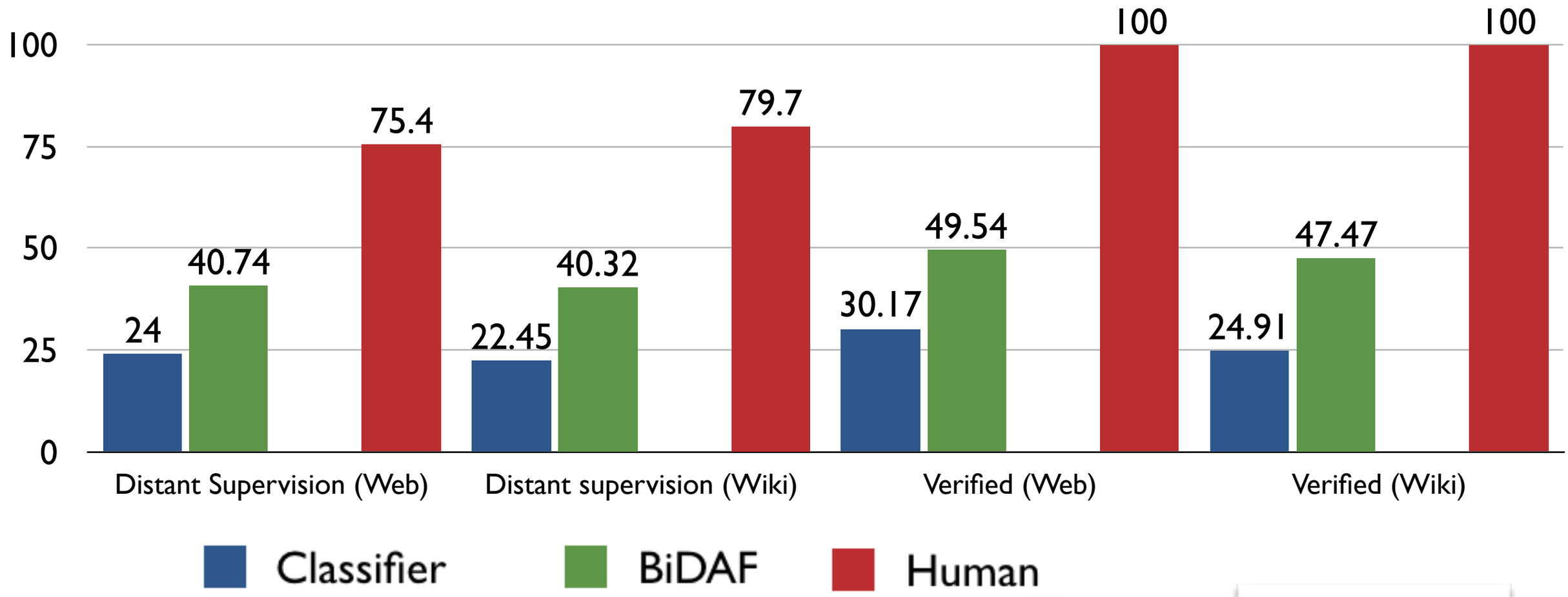
Results (Exact Match)

Distant supervision assumption works for 75-80 % of examples.



Results (Exact Match)

Existing models for SQuAD dataset does not generalize easily.

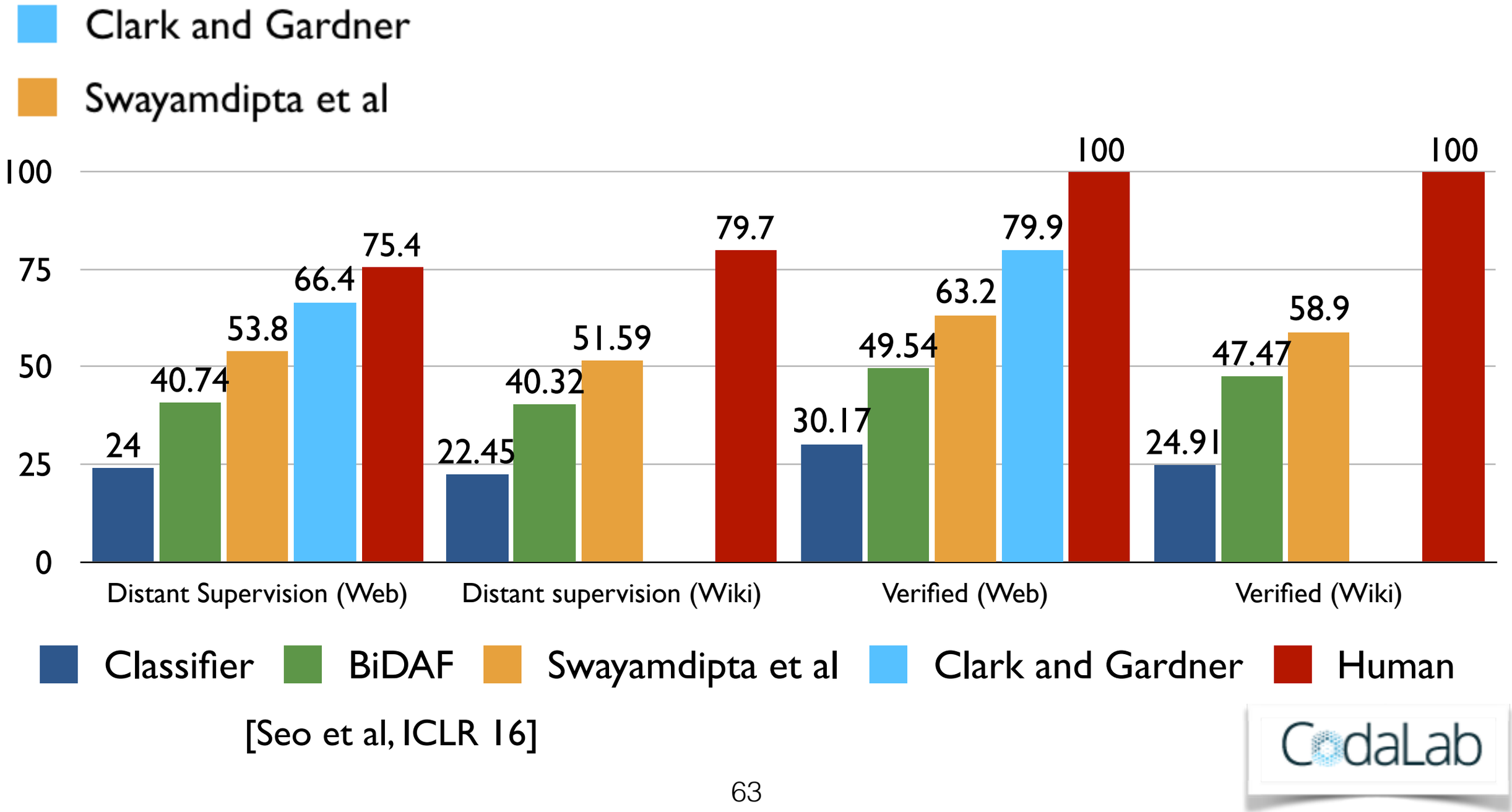


[Seo et al, ICLR 16]

CodaLab

Newer Results

Aggregate over all mentions of the same entity, allow access to longer context.



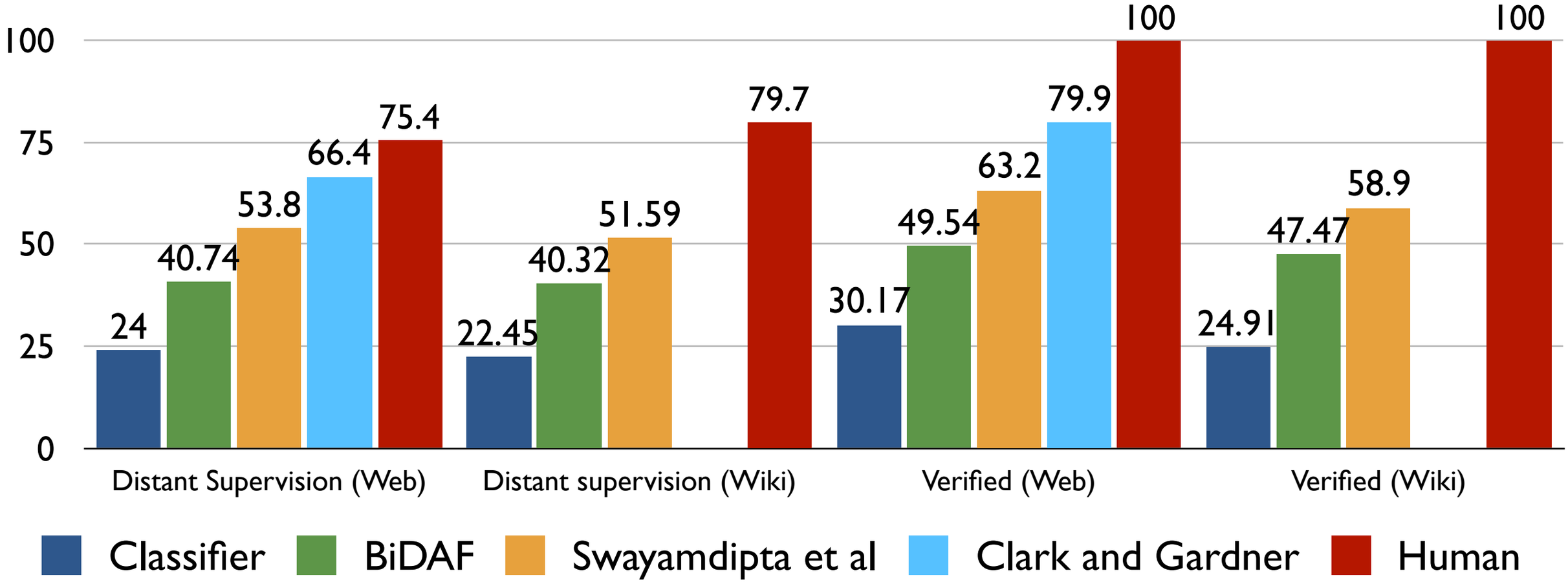
[Seo et al, ICLR 16]

Newer Results

Aggregate over all mentions of the same entity, allow access to longer context.

Clark and Gardner — TF-IDF based sampling of paragraph and shared normalization

Swayamdipta et al — Feedforward network instead of RNN for efficiency, multiple modules.



[Seo et al, ICLR 16]

CodaLab

Improving Model

Introduction

Coarse-to-Fine Question Answering For Long Document [Choi et al, ACL 17]

Improving Data

TriviaQA: Challenge Dataset for Reading Comprehension [Joshi et al, ACL 17]

Applying Model

Reading Comprehension for Relation Extraction [Levy et al, CoNLL17]

Future Work

Zero-Shot Relation Extraction via Reading Comprehension

slides by Omer Levy

Omer Levy, Minjoon Seo,
Eunsol Choi, Luke Zettlemoyer

CoNLL 2017

UWNLP

Question Answering from Raw Text

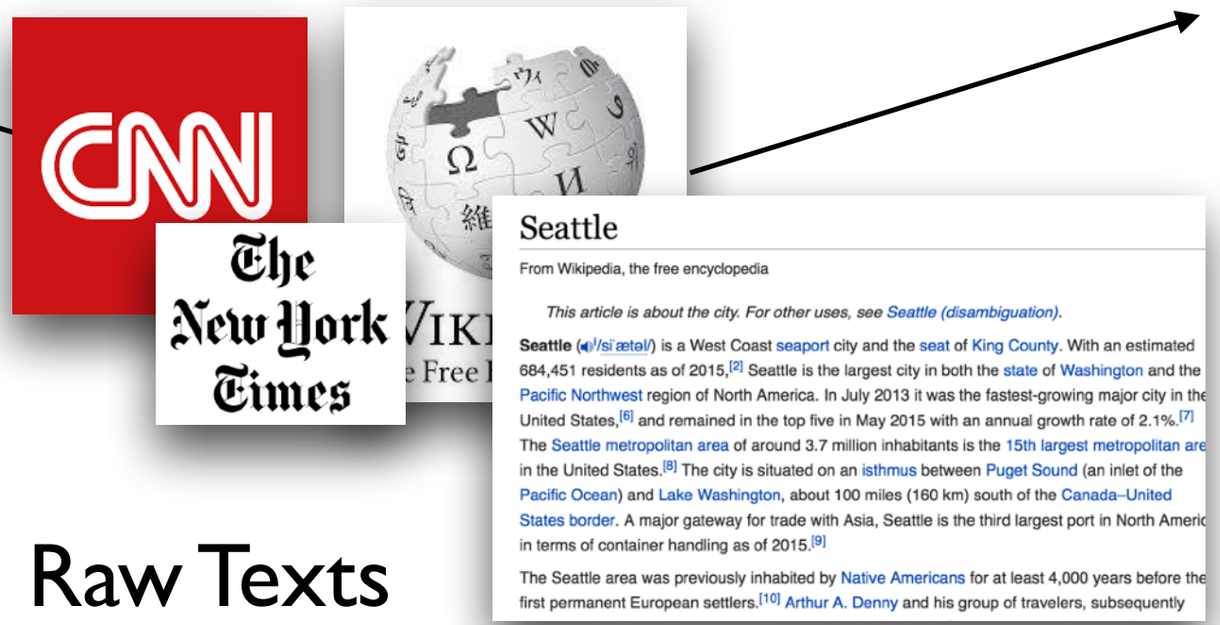
Reading Comprehension

Related Dataset:

WikiQA (Yang et al 15)
CNN dataset (Hermann et al 14)
Children Book Test (Hill et al 15)
SQUAD (Rajpurkar et al 16)
Trivia QA (Joshi et al, 17)

Query

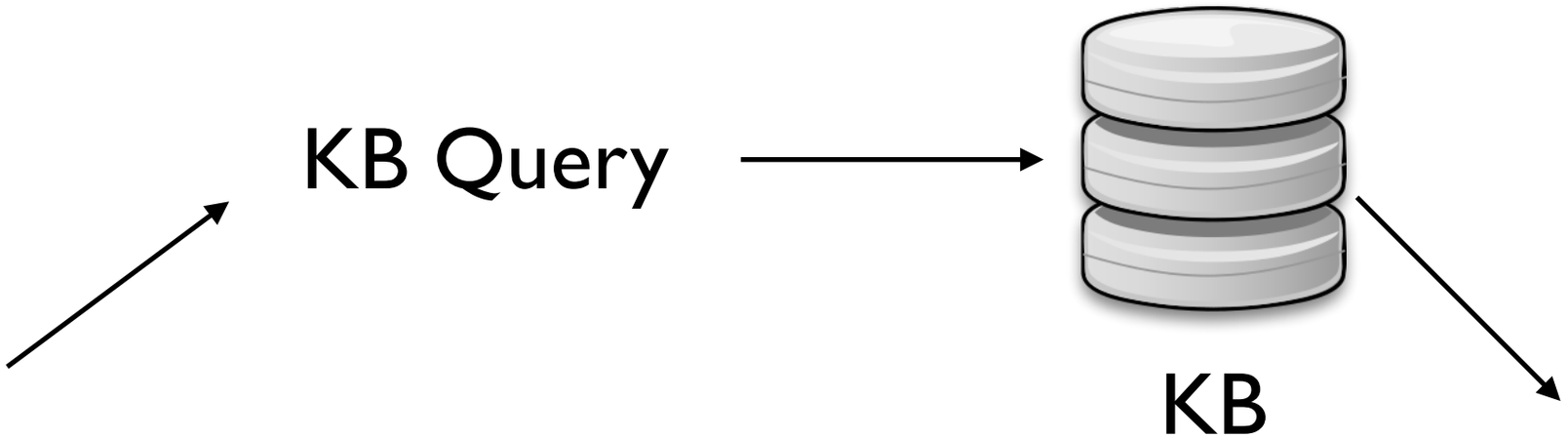
Answer



Raw Texts

Question Answering from KB

Semantic Parsing



Various, Larger Scale KB:

WIKIDATA

Freebase

yago
select knowledge

NELL

Related Work:
[Wong & Mooney 2007],
[Zettlemoyer & Collins 2005, 2007],
[Kwiatkowski et.al 2010, 2011],
[Liang et.al. 2011], [Cai & Yates 2013],
[Berant et.al. 2013],
[Kwiatkowski et.al. 2013], [Yih et al, 15]
[Reddy et.al, 2014], [Wang et al, 15]

Two Sources of Information

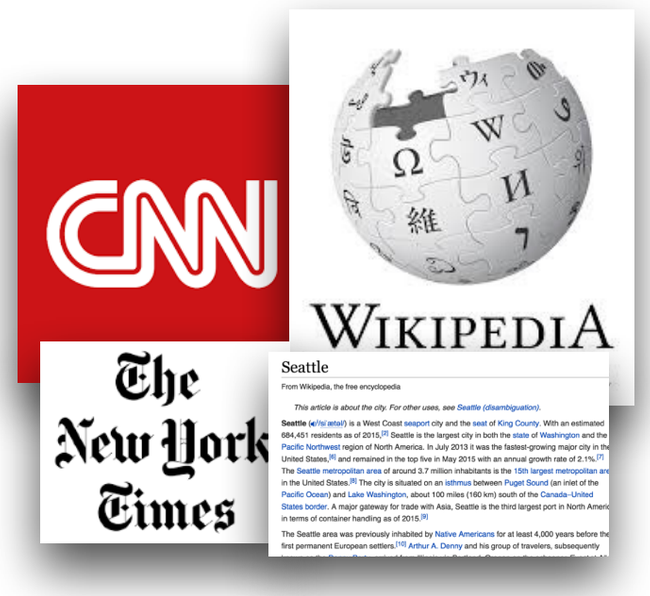
KnowledgeBase

- Can handle compositional logical forms better
  - Count: How many children does Jerry Seinfeld has?

$$\lambda x.\mathrm{eq}(x, \mathtt{count}(\lambda y.\mathtt{person.children}(\mathtt{jerry\_seinfeld}, y)))$$

- ***Efficient Inference***

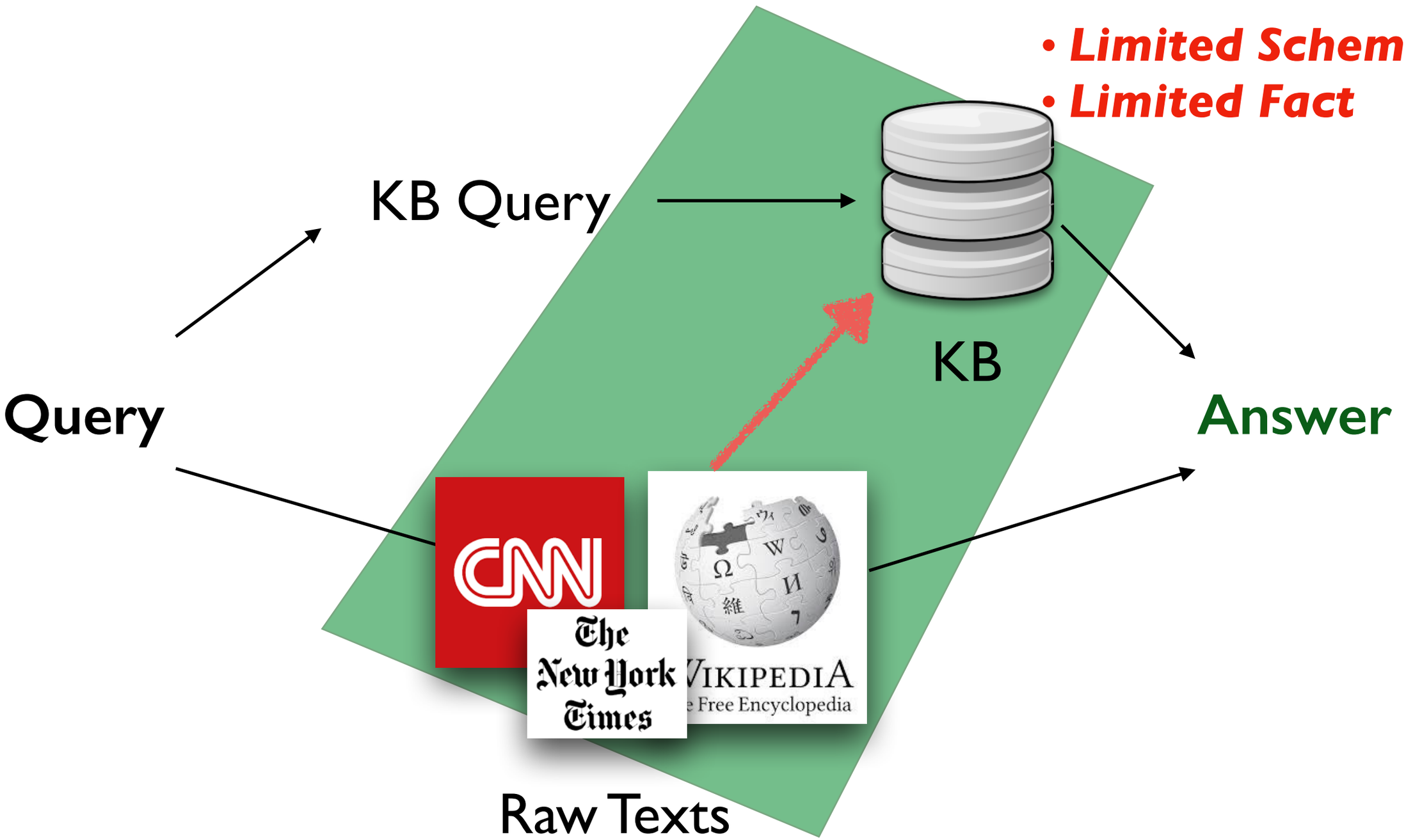
Raw Texts



- ***Contains much more information***
  - Knowledgebase is hugely incomplete.
  - 93% of questions pruned as Freebase could not answer (WebQuestions)

Knowledge Base Population

- ***Efficient Inference***
- ***Limited Schema***
- ***Limited Fact***



Knowledge Base Population

- ***Efficient Inference***
- ***Limited Schema***

Can we use recent advances in reading comprehension models to populate KBs?

Query

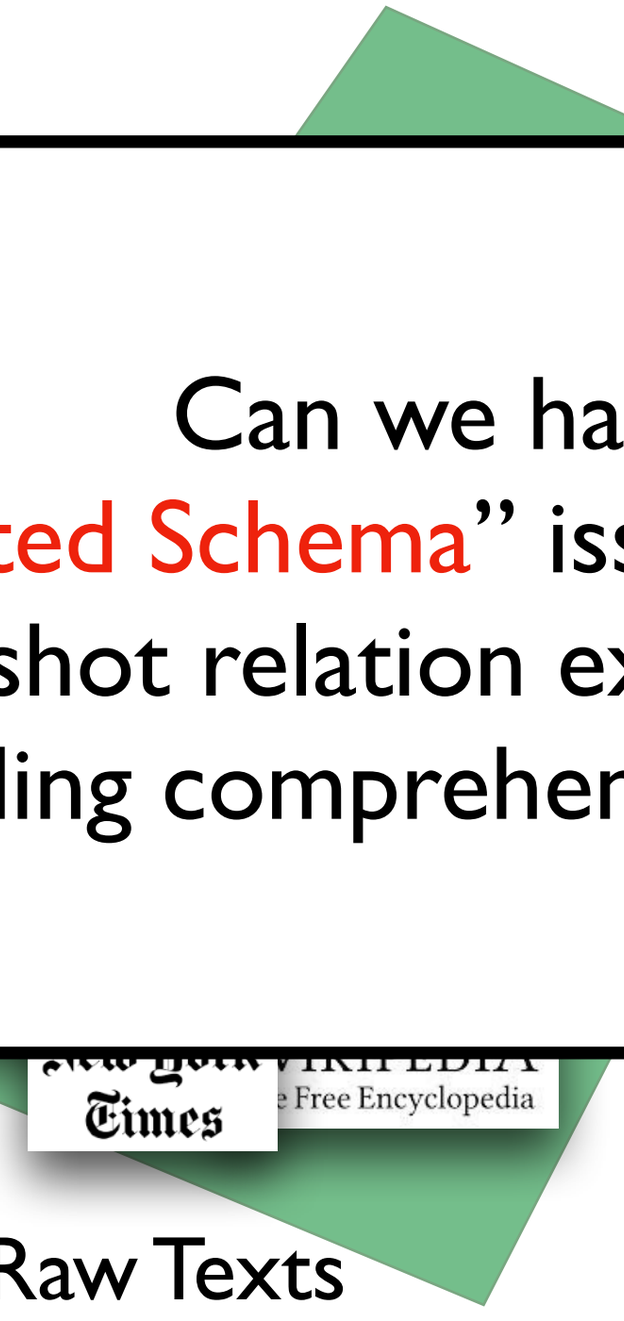
Raw Texts

Knowledge Base Population

- ***Efficient Inference***
- ***Limited Schema***

Can we handle "Limited Schema" issue by doing a zero shot relation extraction with reading comprehension model?

Query



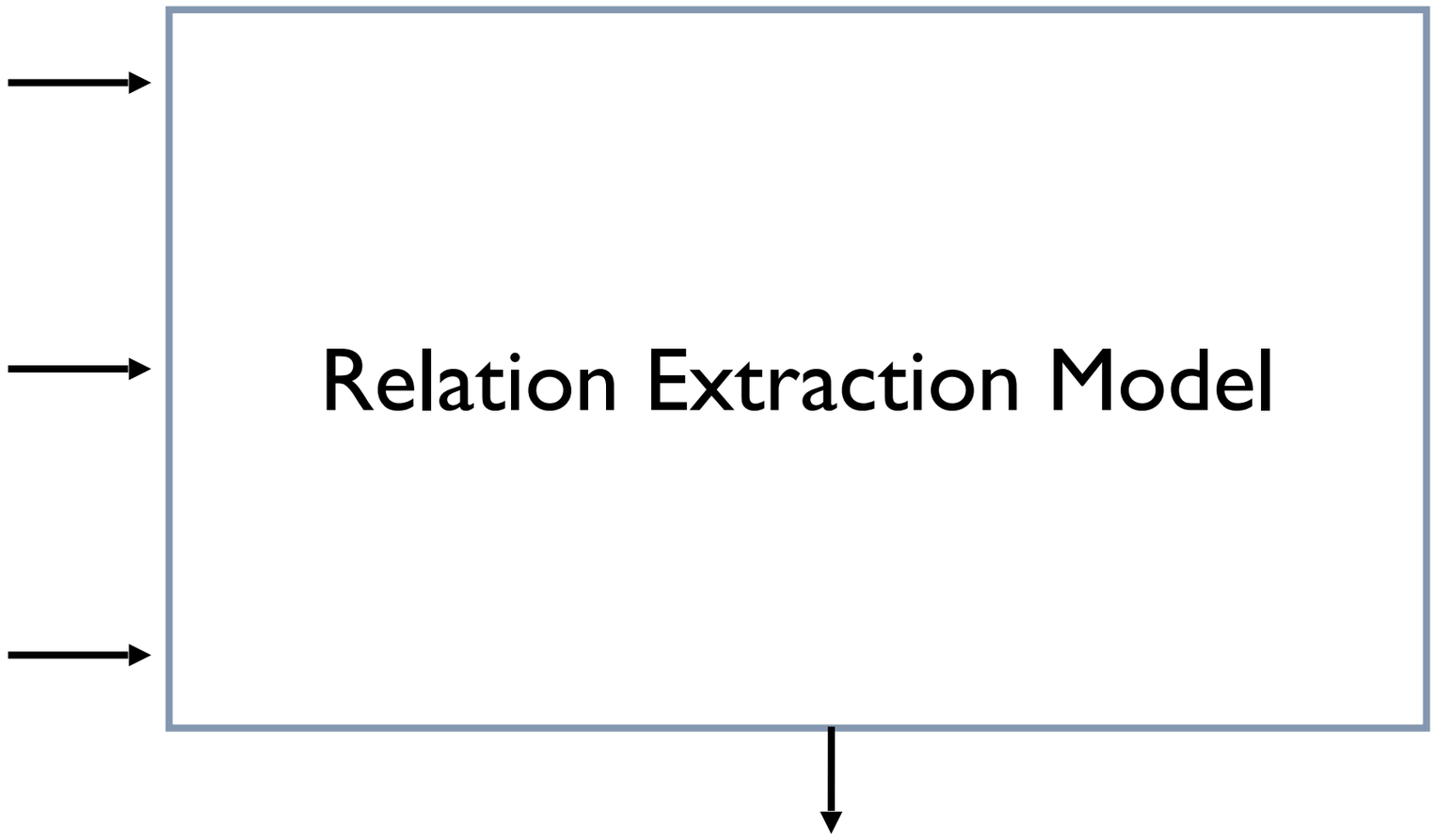
Raw Texts

Relation Extraction (Slot Filling)

Relation:
educated_at(x,?)

Entity: x = Turing

Sentence: "Alan Turing graduated from Princeton."



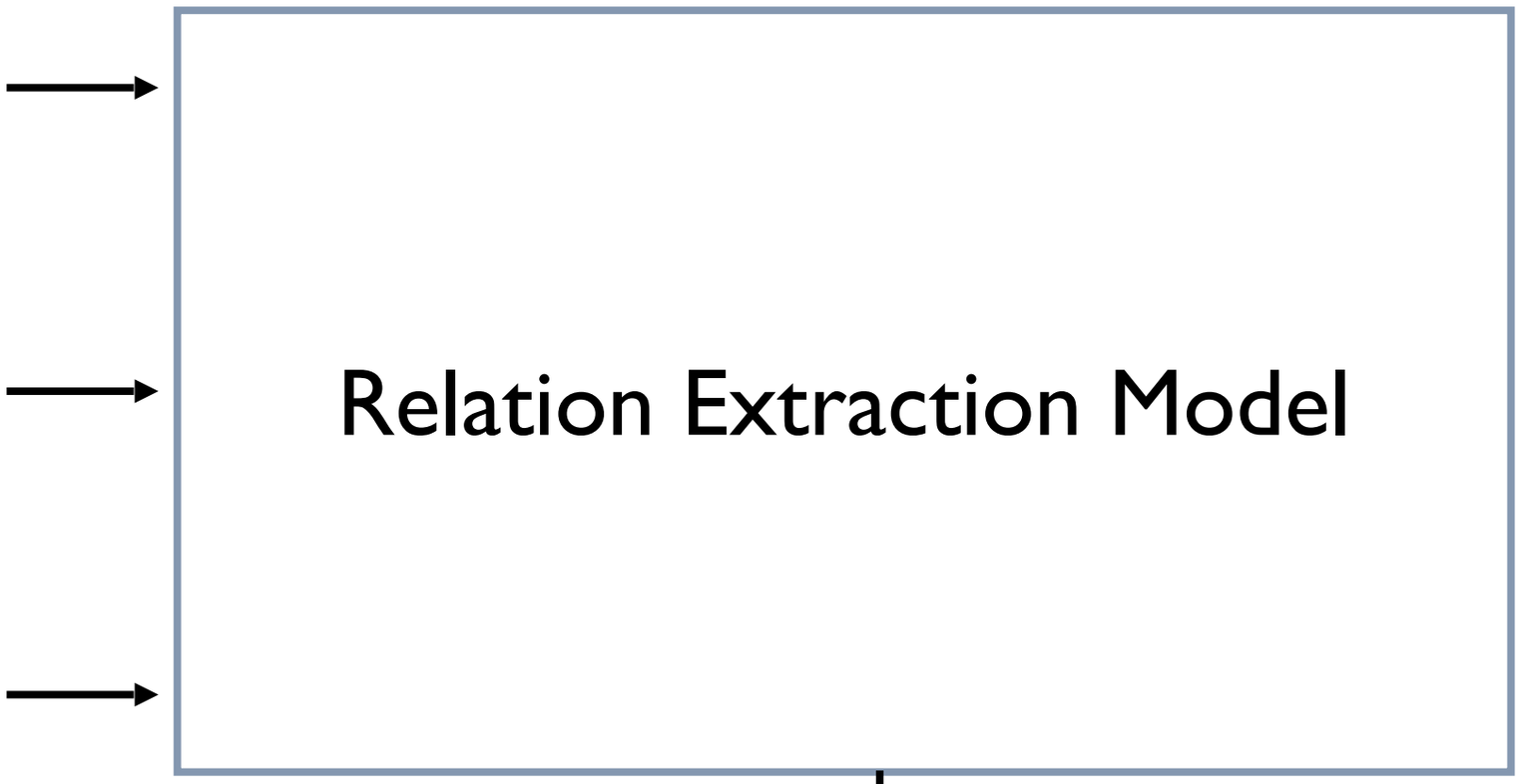
Answer:
Princeton

Relation Extraction (Slot Filling)

Relation:
educated_at(x,?)

Entity: x = Turing

Sentence: "Turing was an English mathematician."

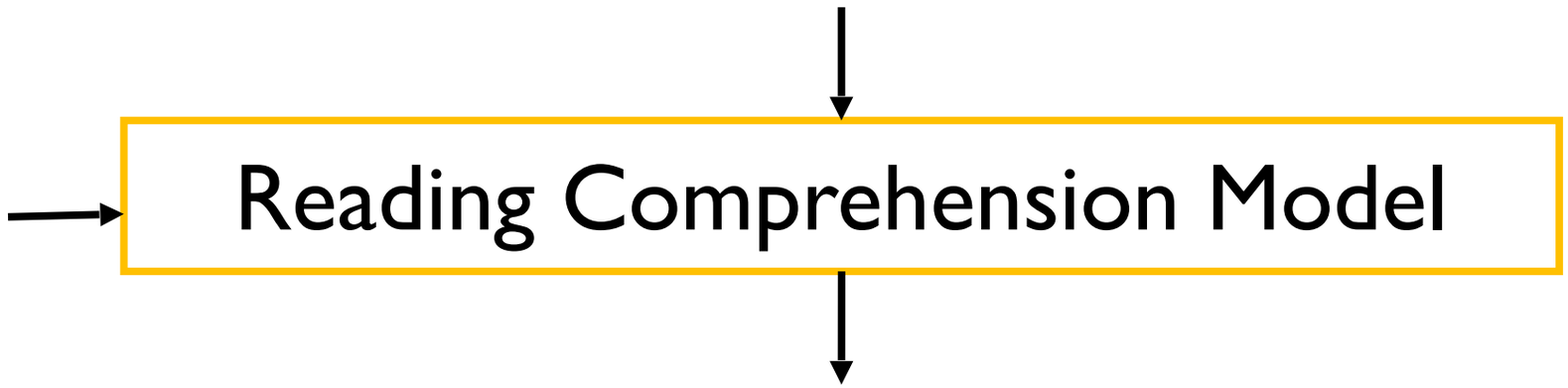


Answer:
<null>

Reading Comprehension

Question:
"Where did Turing study?"

Sentence: "Alan Turing graduated from Princeton."



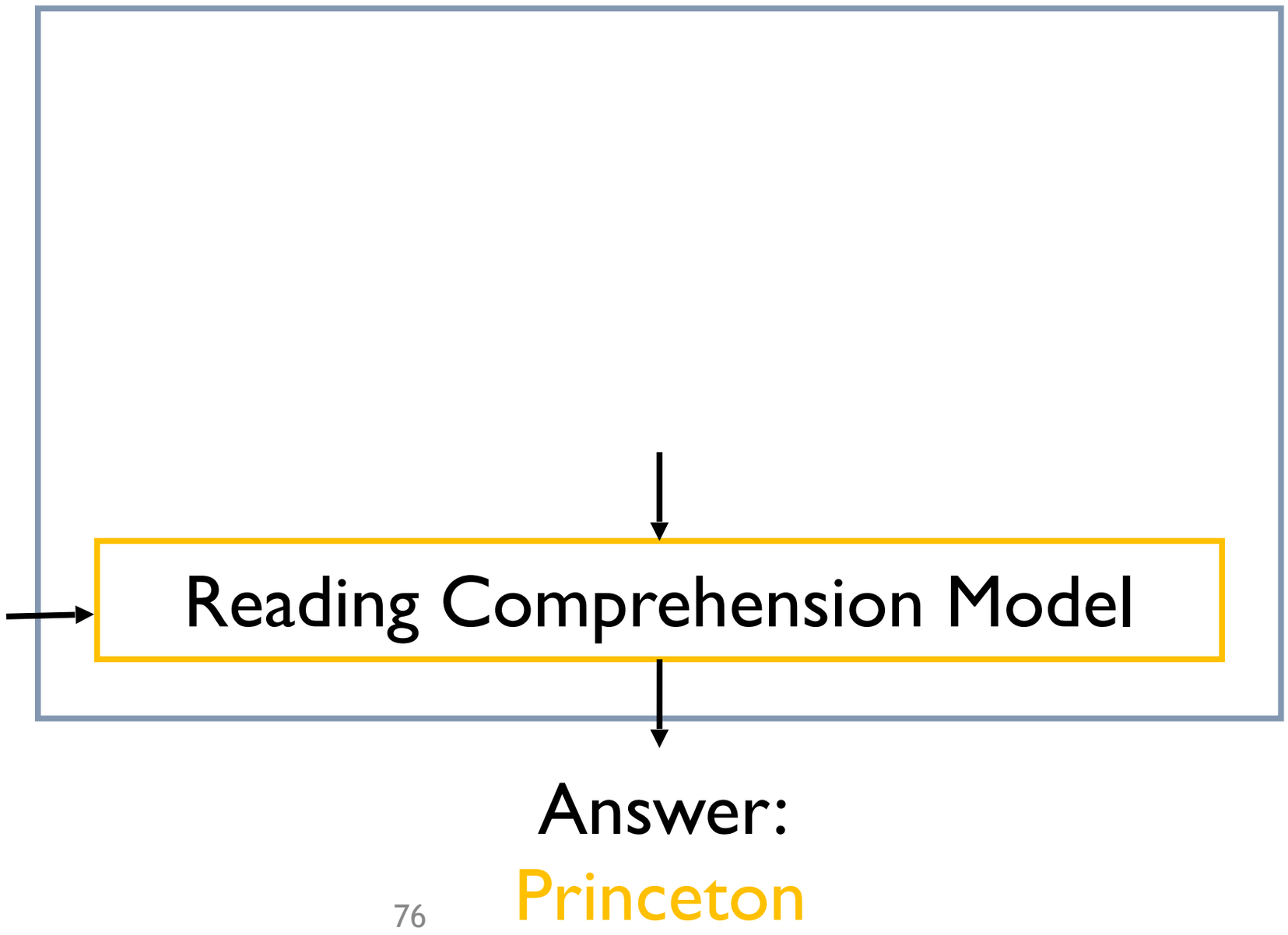
Answer:
Princeton

Relation Extraction via Reading Comprehension

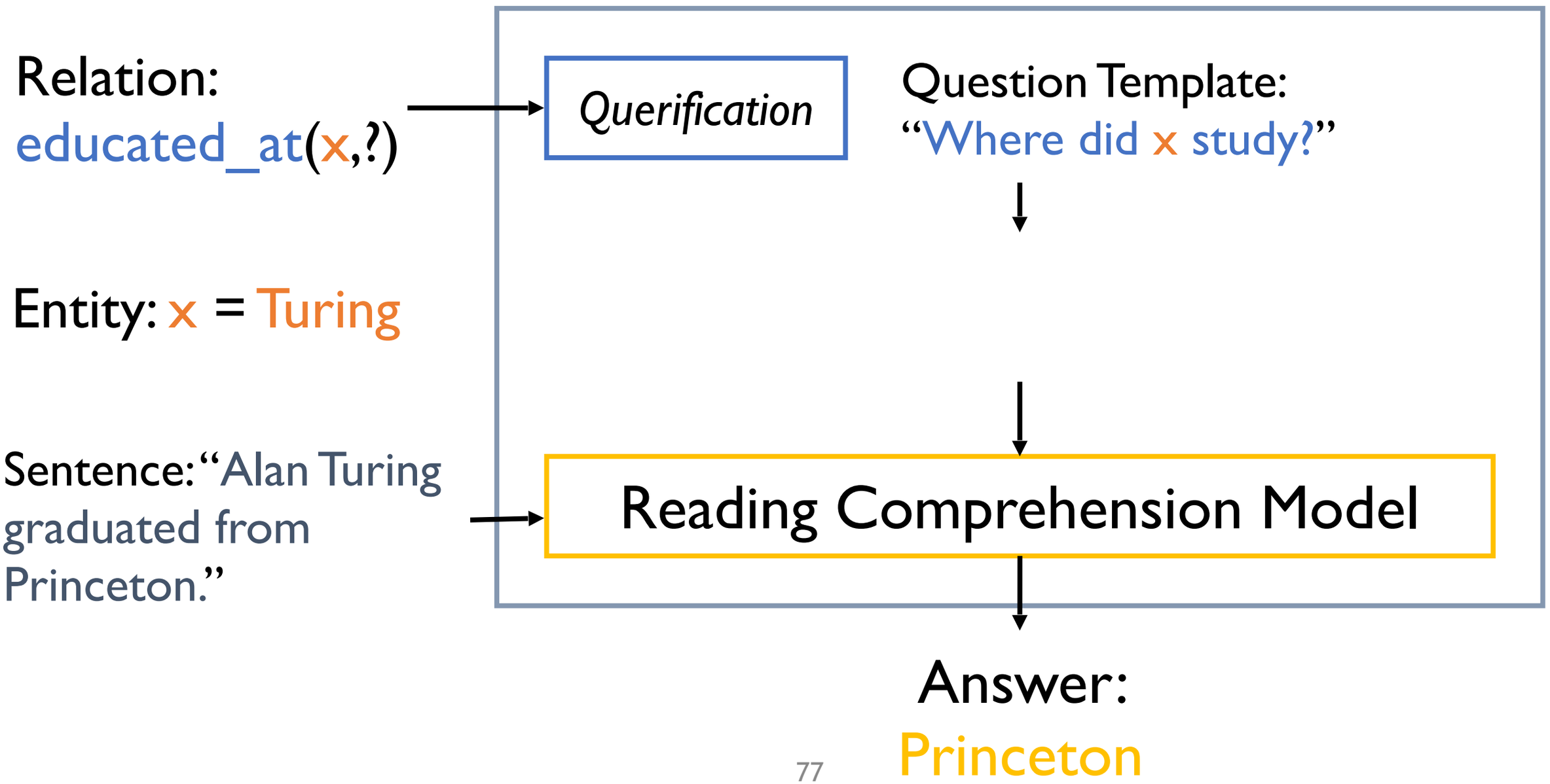
Relation:
educated_at(x,?)

Entity: x = Turing

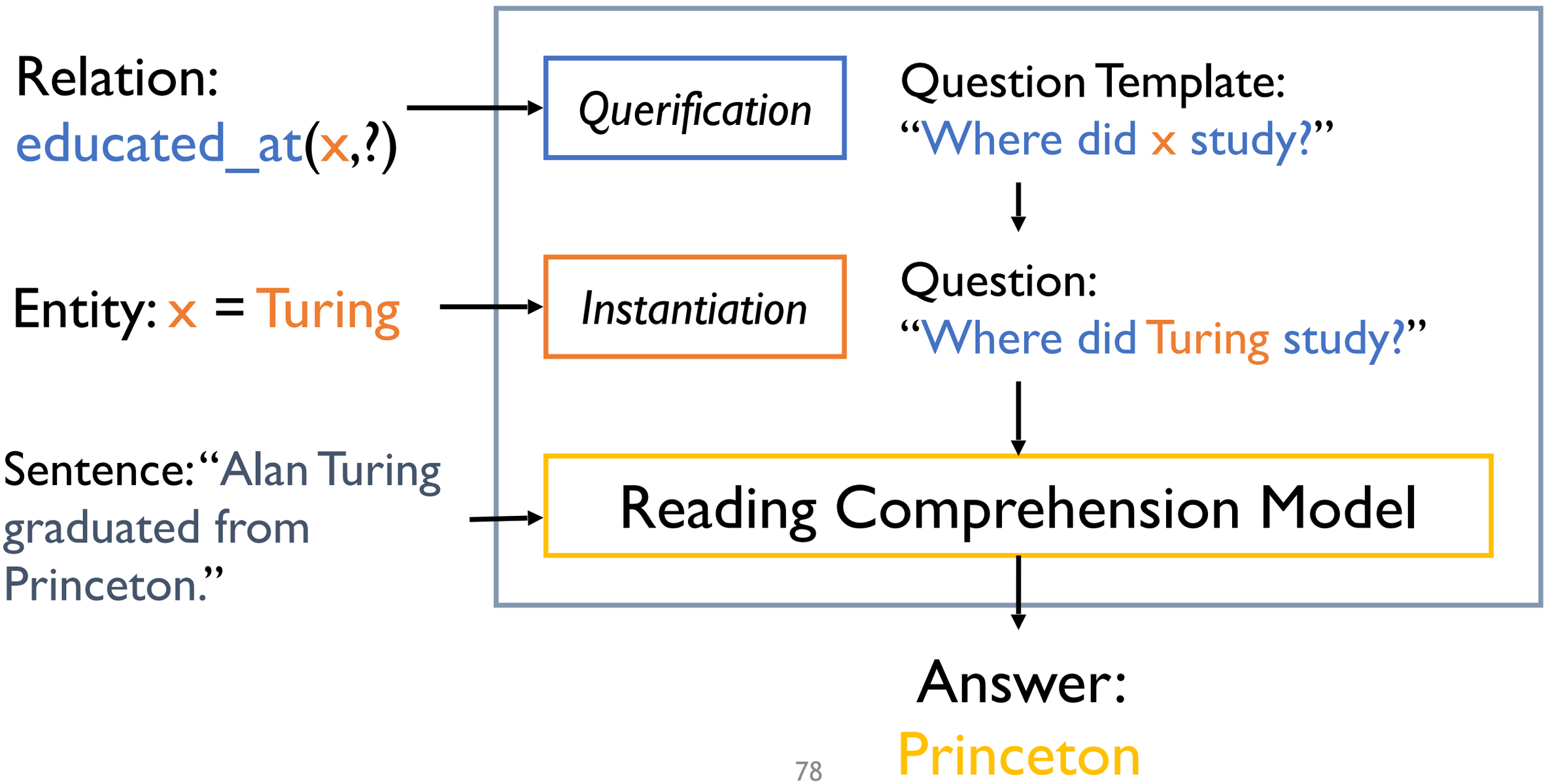
Sentence: "Alan Turing graduated from Princeton."



Relation Extraction via Reading Comprehension



Relation Extraction via Reading Comprehension



Dataset

- Annotated 120 relations from WikiReading (Hewlett et al, ACL 2016)
- Collected 10 templates per relation with high agreement
- Generated over 30,000,000 reading comprehension examples
- Generated negative examples by mixing questions about same entity

Generalizing to Unseen Questions

- Experiment: split the data by question templates

- Performance on seen question templates: 86.6% F1
- Performance on unseen question templates: 83.1% F1

- Our method is robust to new descriptions of existing relations

Generalizing to Unseen Relations

- Model is trained on several relations

  "Where did Alan Turing study?" (educated_at)

  "What is Ivanka Trump's job?" (occupation)

  "Who is Justin Trudeau married to?" (spouse)

- User asks about a new, unseen relation

  "In which country is Seattle located?" (country)

Generalizing to Unseen Relations

- **Experiment:** split the data by **relations**

Results

- Random named-entity baseline: 12.2% F1
- Off-the-shelf RE system: *impossible*
- BiDAF w/ relation name as query: 33.4% F1
- BiDAF w/ querified relation as query: 39.6% F1
  + multiple questions at test: 41.1% F1

Why does a reading comprehension model enable zero-shot relation extraction?

- It can learn answer types that are used across relations

  Q: When was the Snow Hawk released?

  S: The Snow Hawk is a 1925 film…

- It can detect paraphrases of relations

  Q: Who started the Furstenberg China Factory?

  S: The Furstenberg China Factory was founded by Johann Georg…

Conclusion

- Existing reading comprehension model can be adapted to solve relation extraction.
- Natural-language API for defining and querying relations
- Enables zero-shot relation extraction
- Challenging dataset: nlp.cs.washington.edu/zeroshot/

Introduction

Improving Model

Coarse-to-Fine Question Answering For Long Document

Improving Data

TriviaQA: Challenge Dataset for Reading Comprehension

Applying Model

Reading Comprehension for Relation Extraction

Future Work

Remaining Challenges 1. Scalability

- Analysis on the efficiency / accuracy trade-off
- More flexible sub-document selection
  - Instead of top 1-2 sentences, flexible number of sentences or paragraphs
- More hierarchy to be considered for more challenging datasets

Document Selection | Paragraph Selection | Sentence Selection

Remaining Challenges 2. Beyond factoid questions

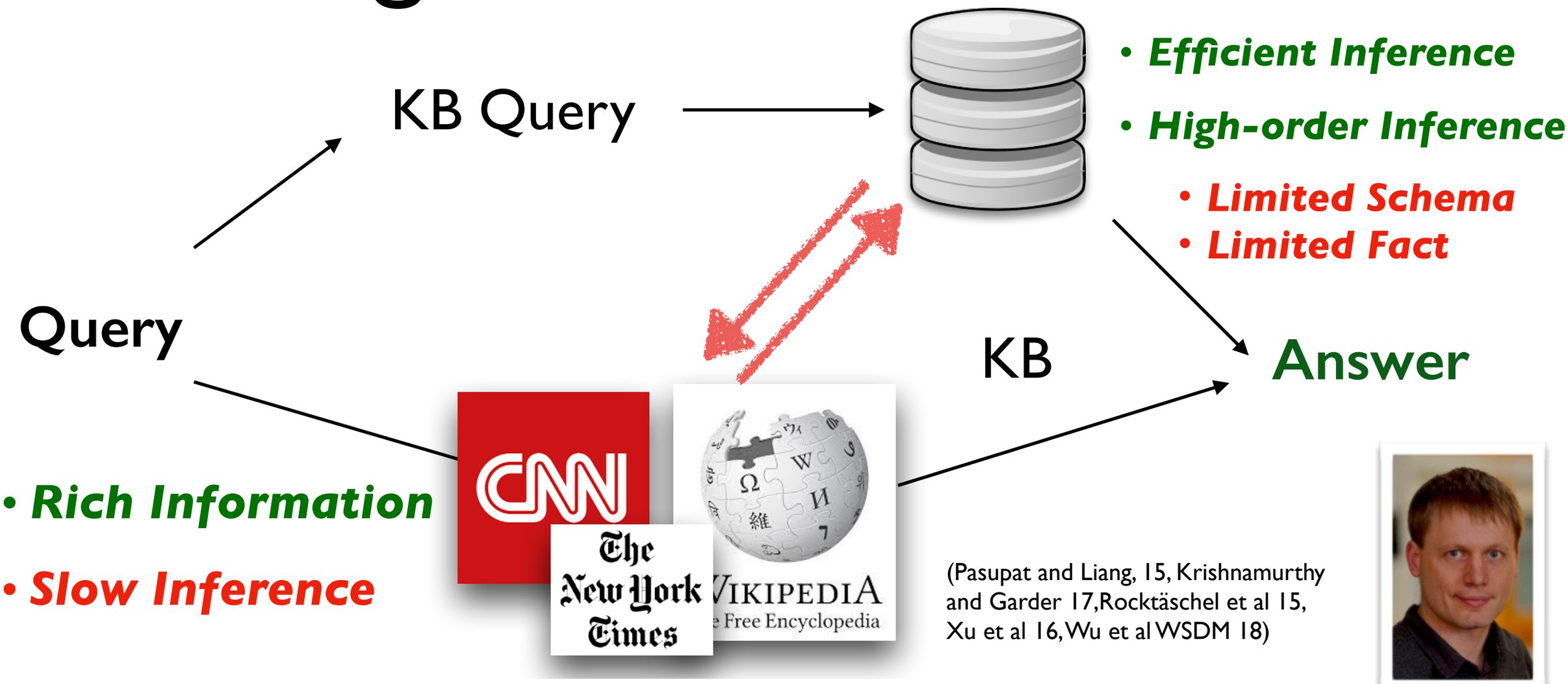
- Questions asking "why".
- Inferred information: Ask most video-game designers about their inspirations. Sam Lake cites Paul Auster's "The Book of Illusions"
  - Sentiment relationship between entities and objects

Movie	Snatch	Revolutionary Road
Question	Why does a robber tell Franky to buy a gun from Boris?	Why does April die?
Story	when you get to London… if you want a gun, call this number.	April dies in the hospital due to complications following the abortion
Answer	Because the robber and Boris want to steal the diamond from Franky	She performs an abortion on h er own.

MovieQA [Tapaswi et al CVPR 16]

Sentiment Graph [Choi et al ACL16]

Remaining Challenges 3. High-Level Inference



How many people have won the Nobel peace prize?

$$\lambda x.\mathtt{eq}(x, \mathtt{count}(\lambda y.\mathtt{person}(y) \wedge \mathtt{won}(y, \mathtt{nobel\_peace\_prize})))$$

Thank you!
Questions?

Luke Zettlemoyer, Yejin Choi,
Dan Weld, Omer Levy,
Minjoon Seo, Mandar Joshi

Daniel Hewlett, Jakob Uszkoreit
Illia Polosukhin, Alexandre Lacoste,
Jonathan Berant