Coarse-to-Fine Question Answering for Long Documents

Eunsol Choi

Daniel Hewlett, Jakob Uszkoreit
Illia Polosukhin, Alexandre Lacoste

Jonathan Berant

UW NLP

Research at Google

TEL AVIV UNIVERSITY
Pursuing the Unknown

ACL 2017

Question Answering

[Rajpurkar et al 16]

paragraph-level

Question Answering

The Victoria and Albert Museum (often abbreviated as the V&A), London, is the world's largest museum of decorative arts and design, housing a permanent collection of over 4.5 million objects. It was founded in 1852 and named after Queen Victoria and Prince Albert. The V&A is located in the Brompton district of the Royal Borough of Kensington and Chelsea, in an area that has become known as "Albertopolis" because of its association with Prince Albert, the Albert Memorial and the major cultural institutions with which he was associated. These include the Natural History Museum, the Science Museum and the Royal Albert Hall. The museum is a non-departmental public body sponsored by the Department for Culture, Media and Sport. Like other national British museums, entrance to the museum has been free since 2001.

where is the Victoria and Albert Museum located?
Ground Truth Answers: The V&A is located in the Brompton district of the Royal Borough of Kensington and Chelsea   London   London   in the Brompton district of the Royal Borough of Kensington and Chelsea

how many permanent objects are located there?
Ground Truth Answers: a permanent collection of over 4.5 million objects.   over 4.5 million objects   over 4.5 million   4.5 million

when was the Victoria and Albert museum founded?
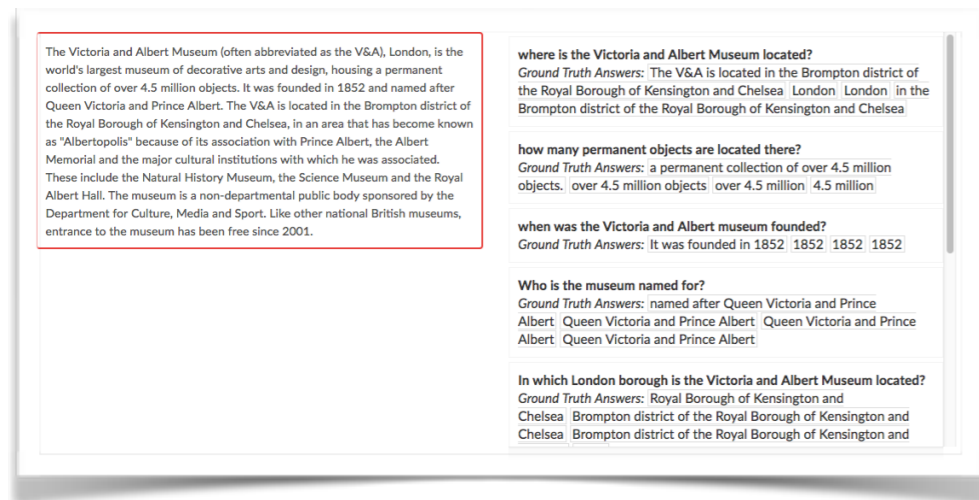*Ground Truth Answers:* It was founded in 1852   1852   1852   1852

Who is the museum named for?
Ground Truth Answers: named after Queen Victoria and Prince Albert   Queen Victoria and Prince Albert   Queen Victoria and Prince Albert   Queen Victoria and Prince Albert

In which London borough is the Victoria and Albert Museum located?
Ground Truth Answers: Royal Borough of Kensington and Chelsea   Brompton district of the Royal Borough of Kensington and Chelsea   Brompton district of the Royal Borough of Kensington and

[Rajpurkar et al 16]

Google

when is fathers day

ALL   NEWS   VIDEOS   IMAGES   MAPS

Sunday, June 18

[TREC QA dataset]

paragraph-level   corpus-level

Question Answering

The Victoria and Albert Museum (often abbreviated as the V&A), London, is the world's largest museum of decorative arts and design, housing a permanent collection of over 4.5 million objects. It was founded in 1852 and named after Queen Victoria and Prince Albert. The V&A is located in the Brompton district of the Royal Borough of Kensington and Chelsea, in an area that has become known as "Albertopolis" because of its association with Prince Albert, the Albert Memorial and the major cultural institutions with which he was associated. These include the Natural History Museum, the Science Museum and the Royal Albert Hall. The museum is a non-departmental public body sponsored by the Department for Culture, Media and Sport. Like other national British museums, entrance to the museum has been free since 2001.

where is the Victoria and Albert Museum located?
Ground Truth Answers: The V&A is located in the Brompton district of the Royal Borough of Kensington and Chelsea London London in the Brompton district of the Royal Borough of Kensington and Chelsea
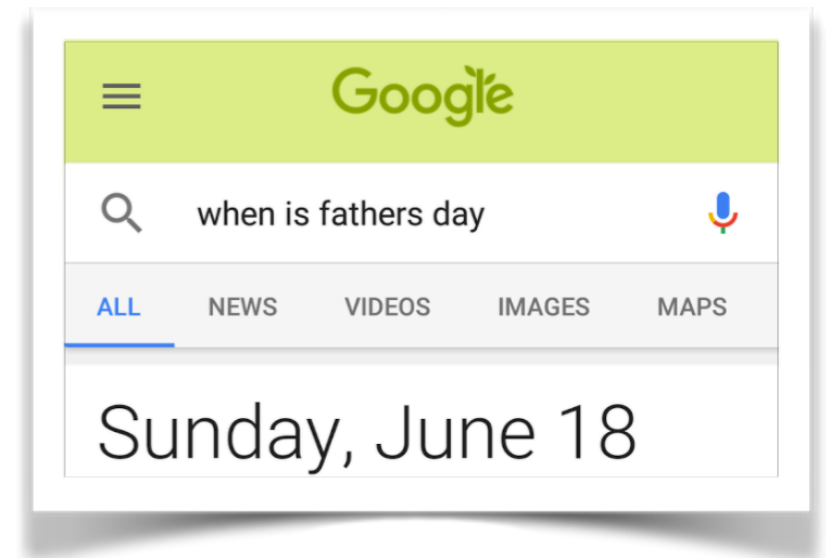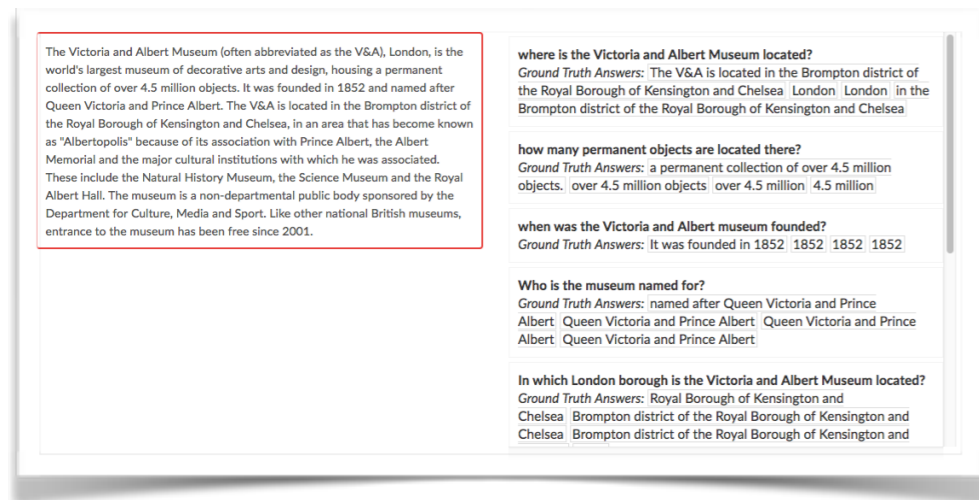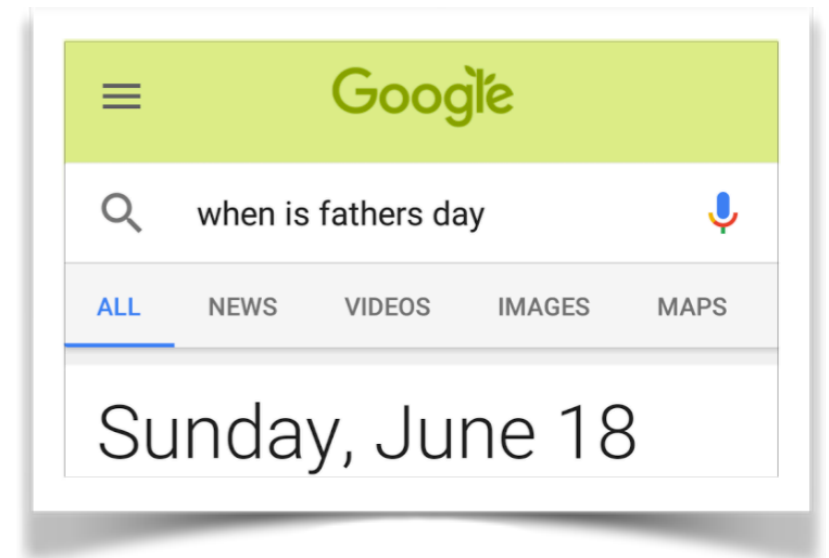
how many permanent objects are located there?
Ground Truth Answers: a permanent collection of over 4.5 million objects. over 4.5 million objects over 4.5 million 4.5 million

when was the Victoria and Albert museum founded?
Ground Truth Answers: It was founded in 1852 1852 1852 1852

Who is the museum named for?
Ground Truth Answers: named after Queen Victoria and Prince Albert Queen Victoria and Prince Albert Queen Victoria and Prince Albert Queen Victoria and Prince Albert

In which London borough is the Victoria and Albert Museum located?
Ground Truth Answers: Royal Borough of Kensington and Chelsea Brompton district of the Royal Borough of Kensington and Chelsea Brompton district of the Royal Borough of Kensington and

Google
when is fathers day
ALL   NEWS   VIDEOS   IMAGES   MAPS

Sunday, June 18

[Rajpurkar et al 16]

[TREC QA dataset]

paragraph-level

corpus-level

Single long document

4

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

How long do effects of lorazepam last?

Question Answering

How long do effects of lorazepam last?

Question Answering

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 benzodiazepine. Lorazepam reduces anxiety, interferes with new memory formation, reduces agitation 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 bo 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

Task

How long do effects of lorazepam last?
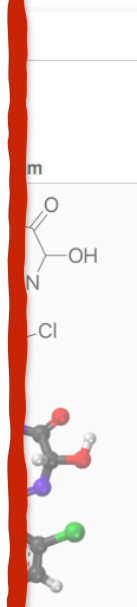
Lorazepam

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

---

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

11

This Work

Query (x)

Document (d)

Sentence Selection

Sentence set $(\hat{s})$

Answer Generation

Answer (y)

This Work

Query (x)

Document (d)

Sentence Selection

Sentence set ($\hat{s}$)

Answer Generation
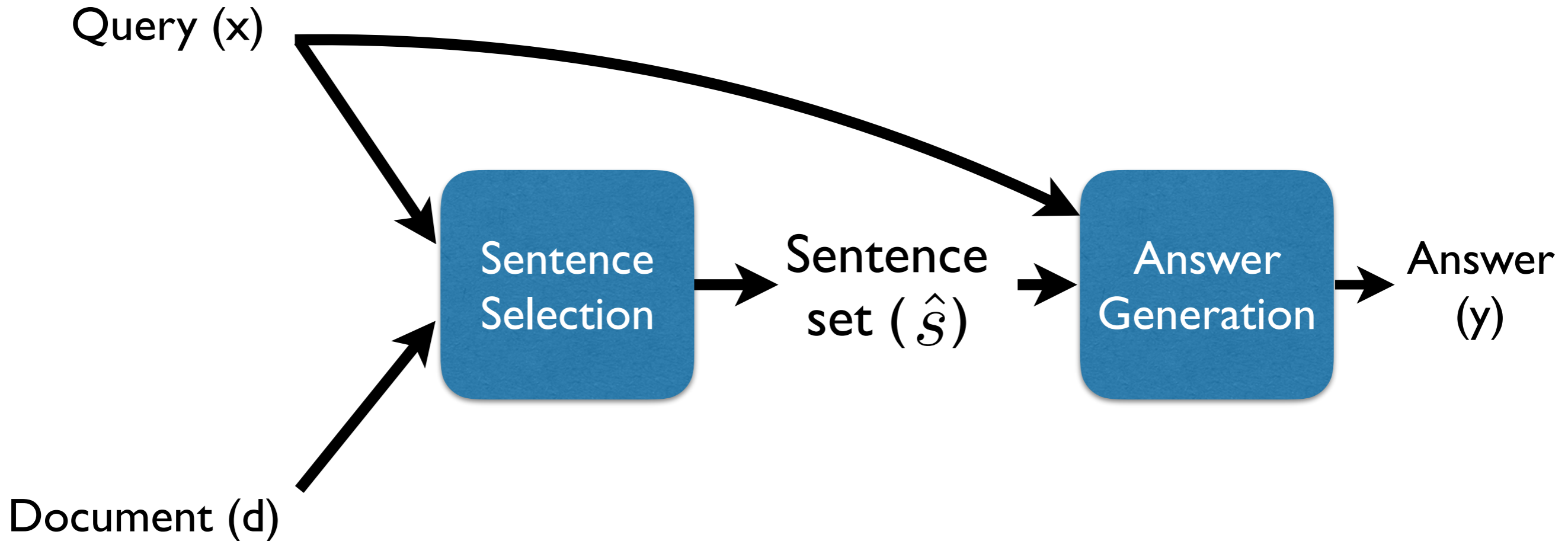
Answer (y)
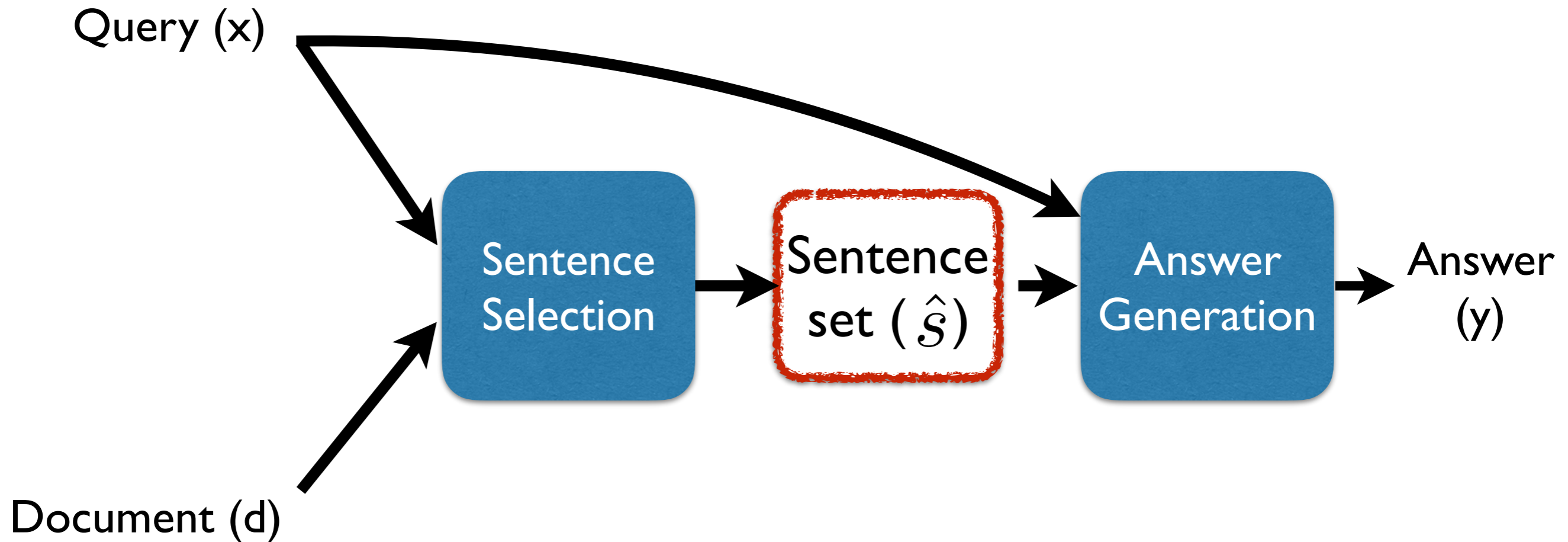
This Work

Query (x)

Document

nswer
(y)

- Coarse-to-Fine model for question answering

- Substantially faster (up to 6.7 times) with comparable accuracies

- Learning without direct supervision for evidence sentence

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)

- 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)

- New datasets with multiple documents
(SearchQA[Dunn et al 17], MS MARCO[Nguyen et al 16], TriviaQA[Joshi et al17])

Data

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

- WikiReading-Long
  - Challenging WikiReading subset, longer documents

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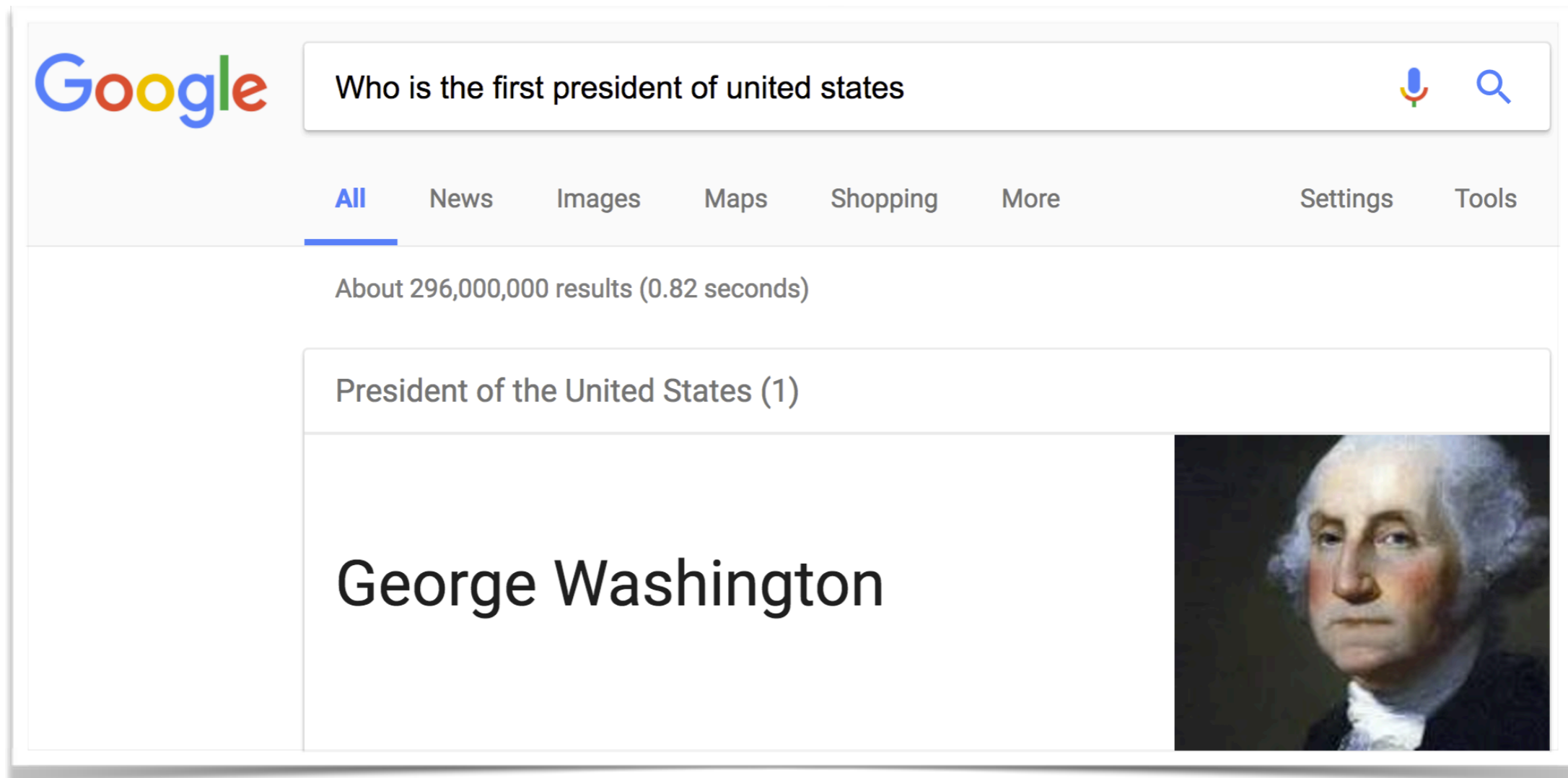
Entity			
Folka			
Canada	to body of water	the Atlantic to the Pacific and northward into the Artic Ocean	ean, Pacific, Arctic Ocean
Breaking Bad	Start time	Breaking Bad is a TV series… from January 20, 2008	20 January 2008

✓ Large-scale dataset covering various domains

✓ Most documents are relatively short (480 words)

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...		
how... gan...		
the...	✓ Large-scale, noisy dataset covering various domain (3.5M)	
wha... pre...	✓ More diverse and natural questions	
ohio...	✓ Including systematically generated noise (~25%)	
how old is ed marinaro		born on March 31, 1950
allers syndrome		Ehlers-Danlos

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)

Document (d)

Sentence Selection → Sentence set ($\hat{s}$) → Answer Generation → Answer (y)

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

25
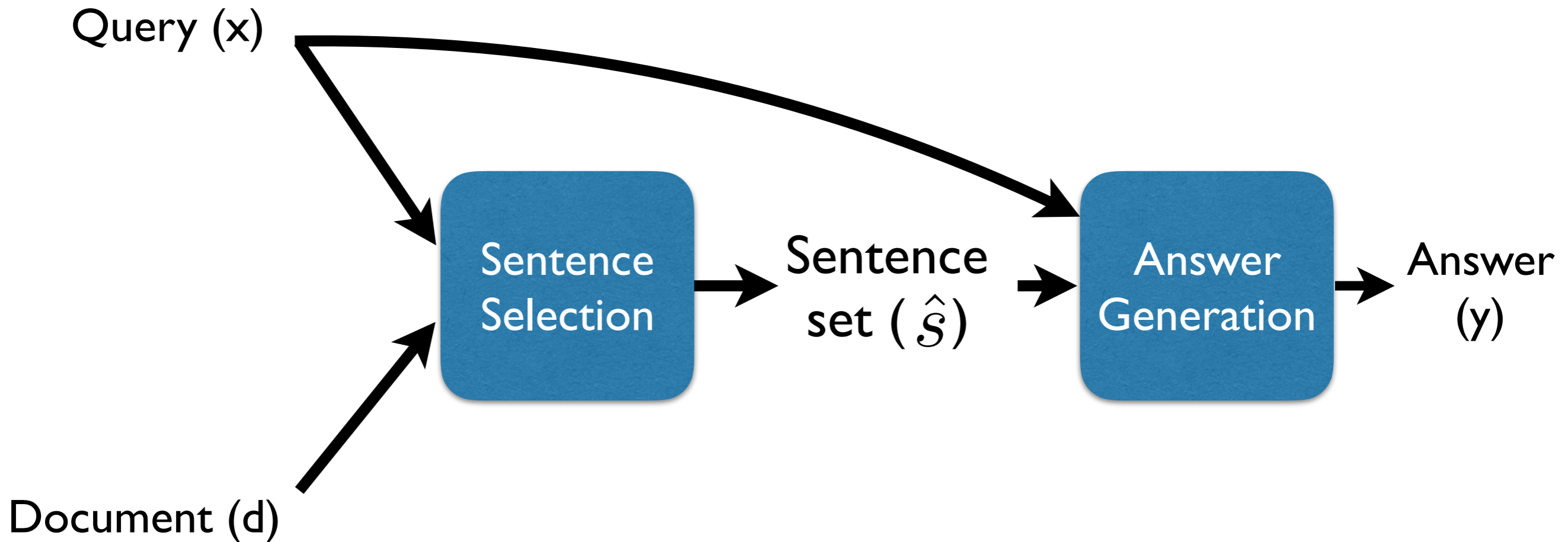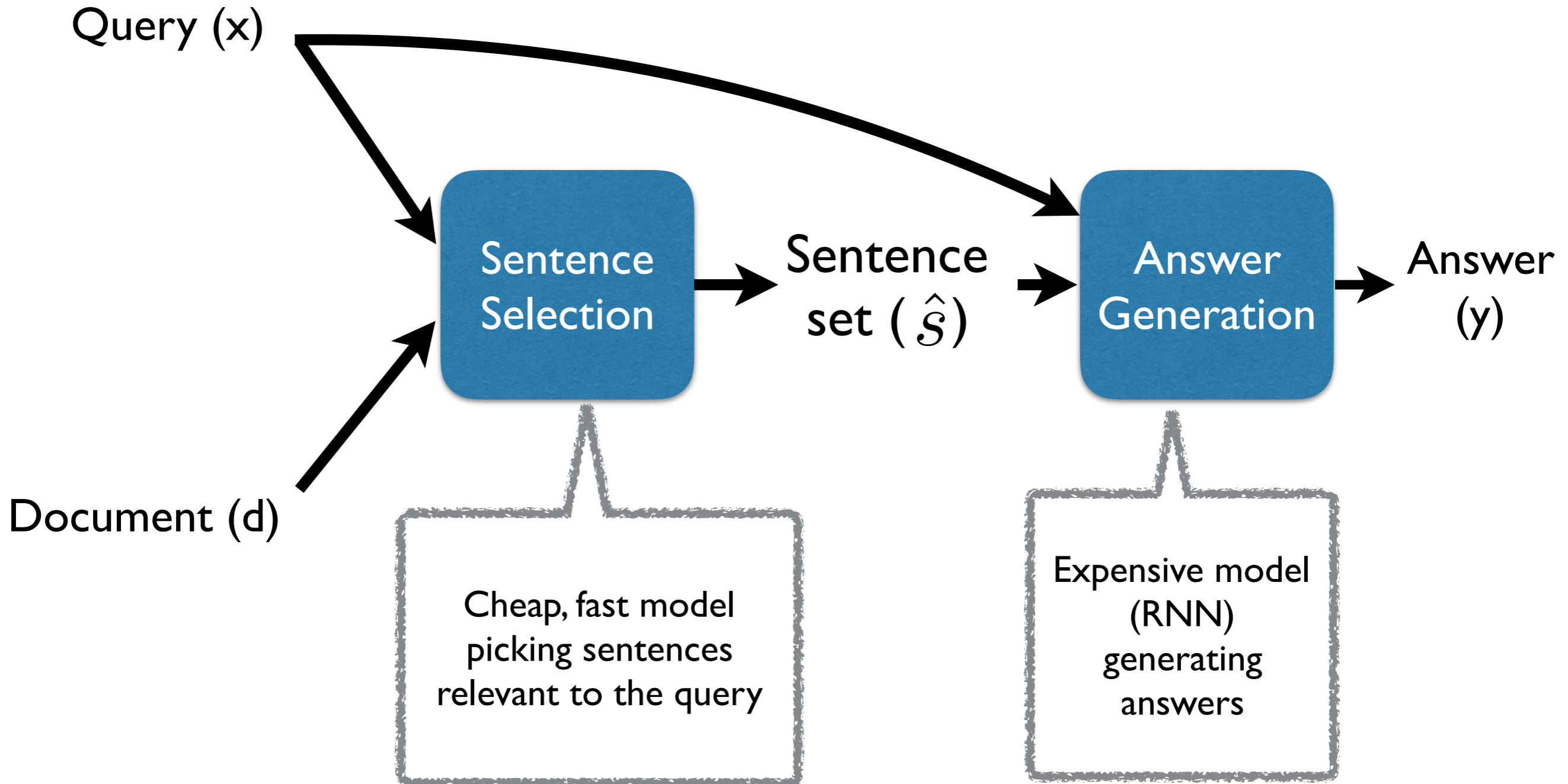
Sentence Selection Model

- Take query and document as input
- Computes relevance score for each sentence $(P(s|x, d))$ to generate sentence set to pass on to answer generation model.
- Coarse and fast sentence representation (BoW)

SoftMax

↑

Hidden Layer

↑
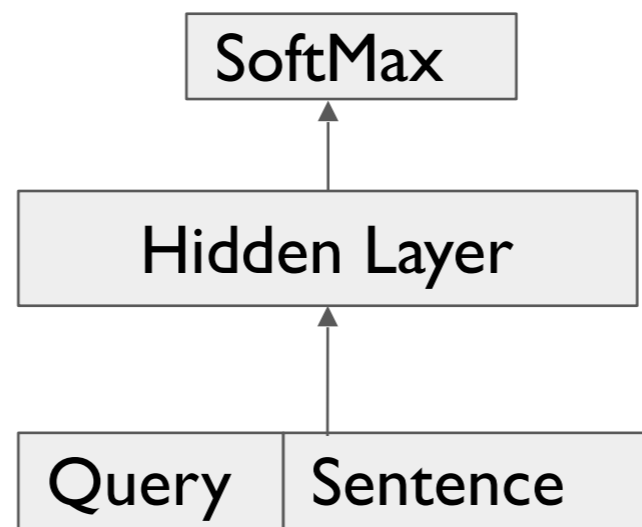
Query	Sentence

Answer Generation Model
(Hewlett et al, ACL16)

- Given a sentence set and a query, generate answer strings

- RNN encoder-decoder model with placeholders

Learning

Query (x)

Document (d)

Sentence Selection → **Sentence set ($\hat{s}$)** → **Answer Generation** → Answer (y)
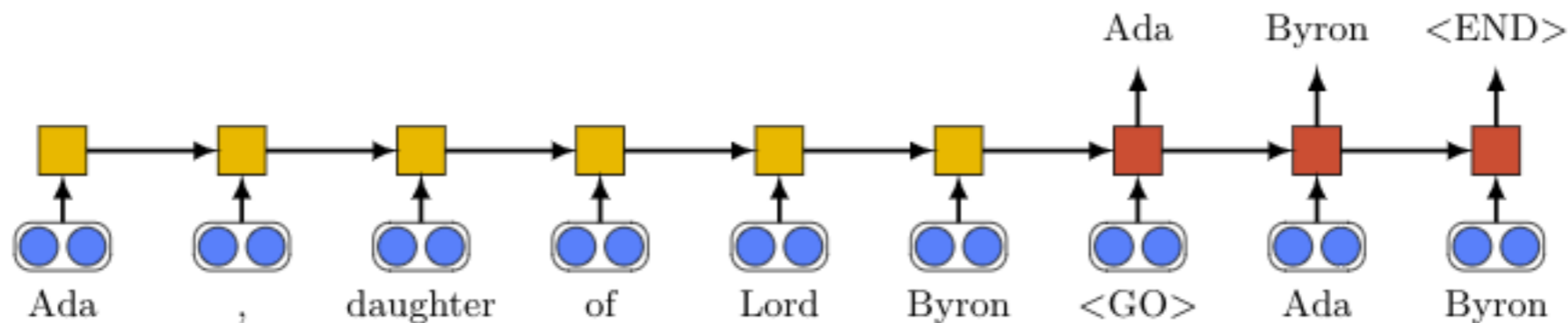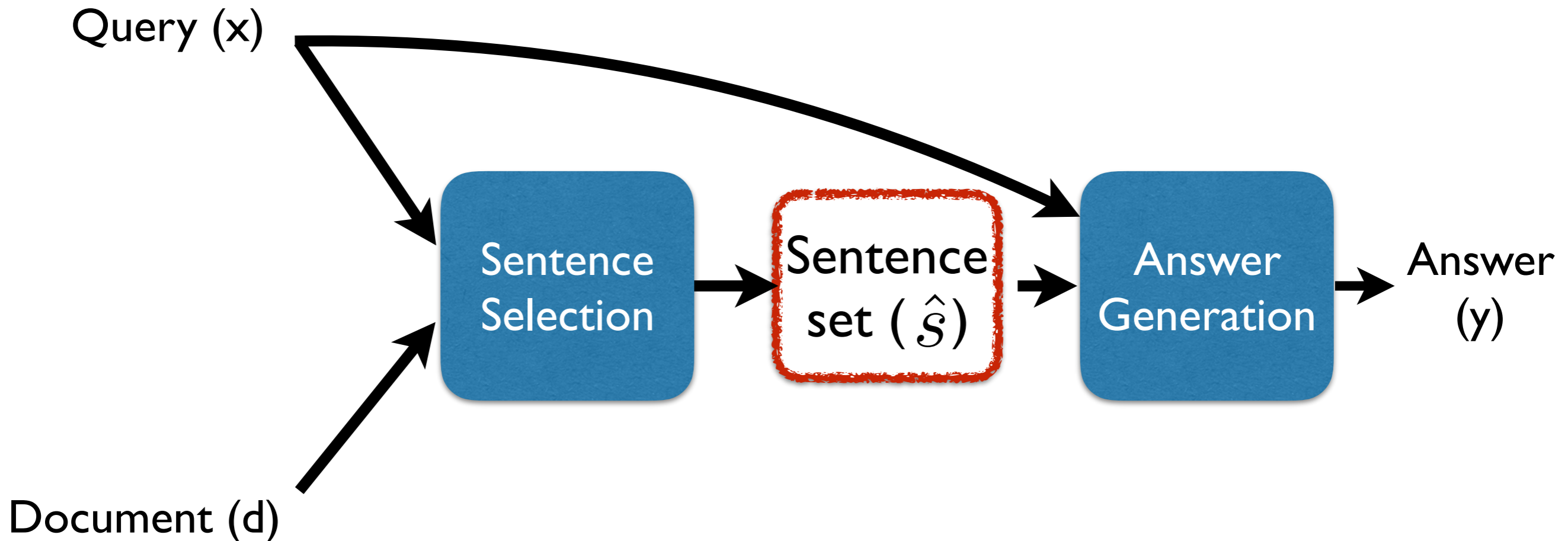
We do not have supervision for which sentences contain the information.

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: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.

False Negative!

False Positive!

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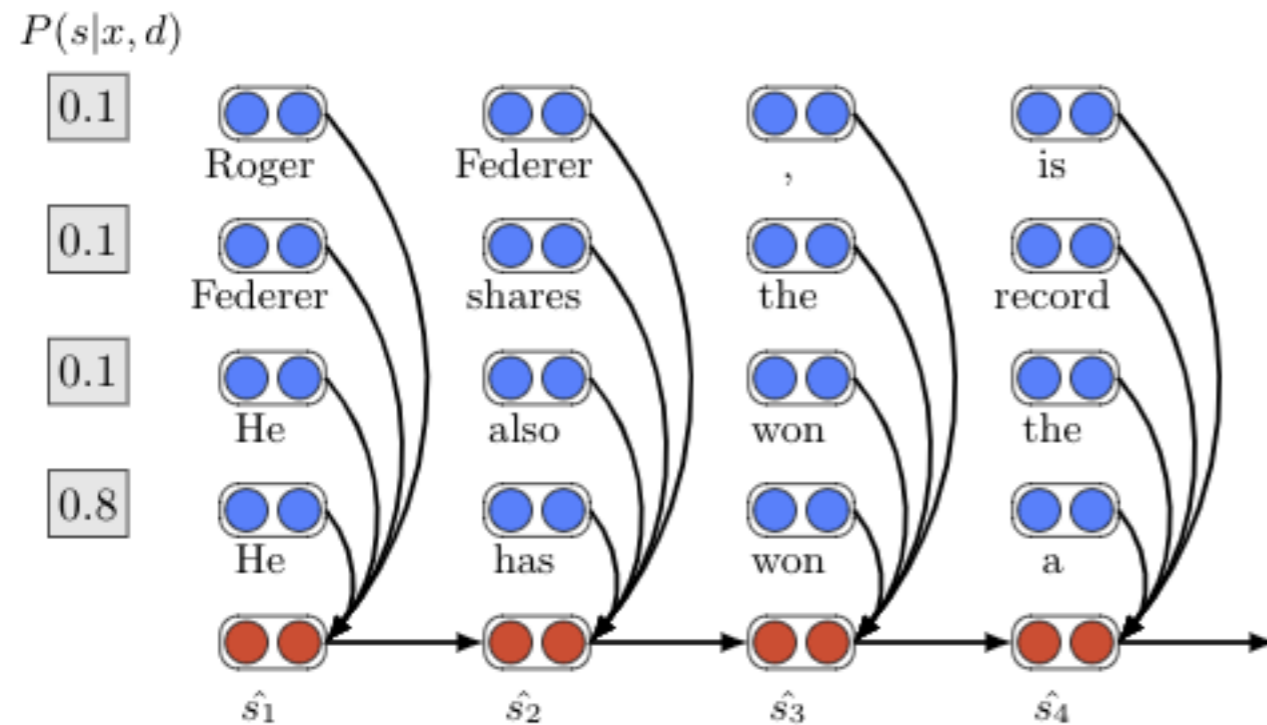
- Make a "blended" token representation by merging each sentence token weighted by its relevancy score p(s|x).

- 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
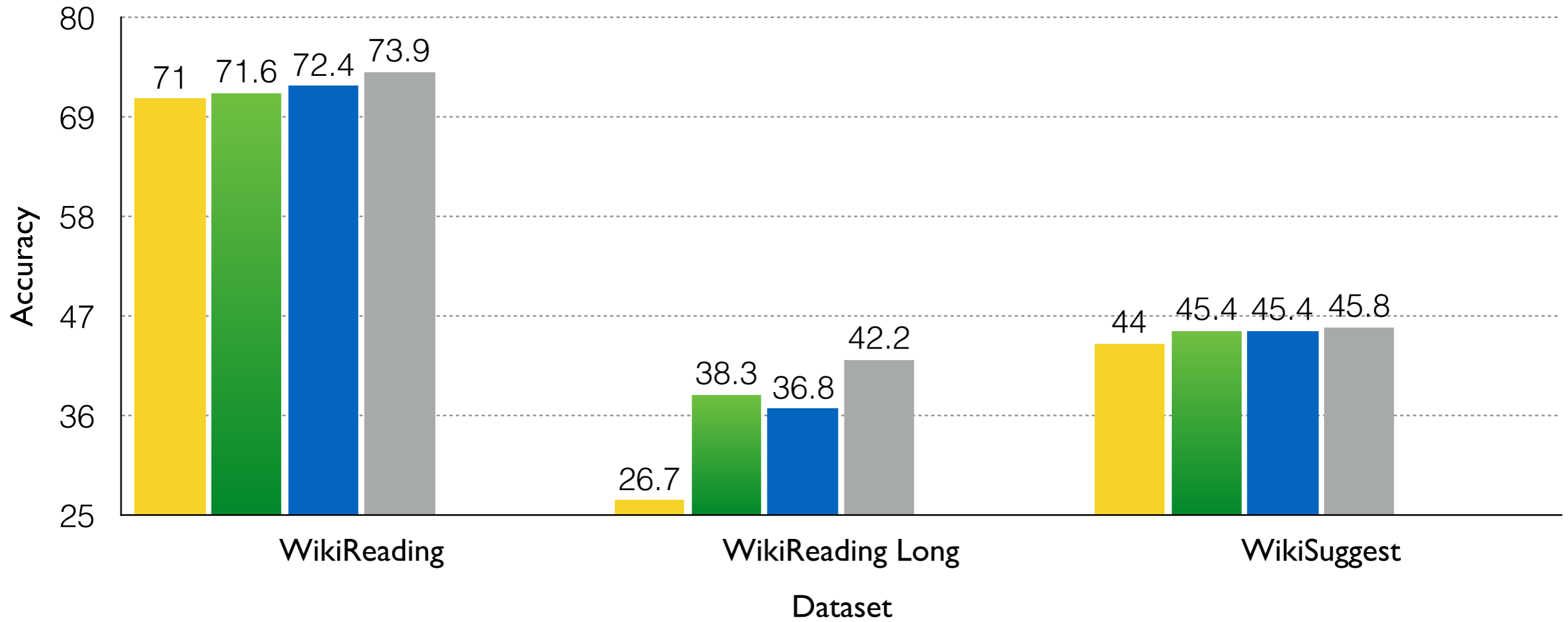
Evaluation

- Answer accuracy:
  - exact match accuracy

- Efficiency:
  - time to finish document encoding
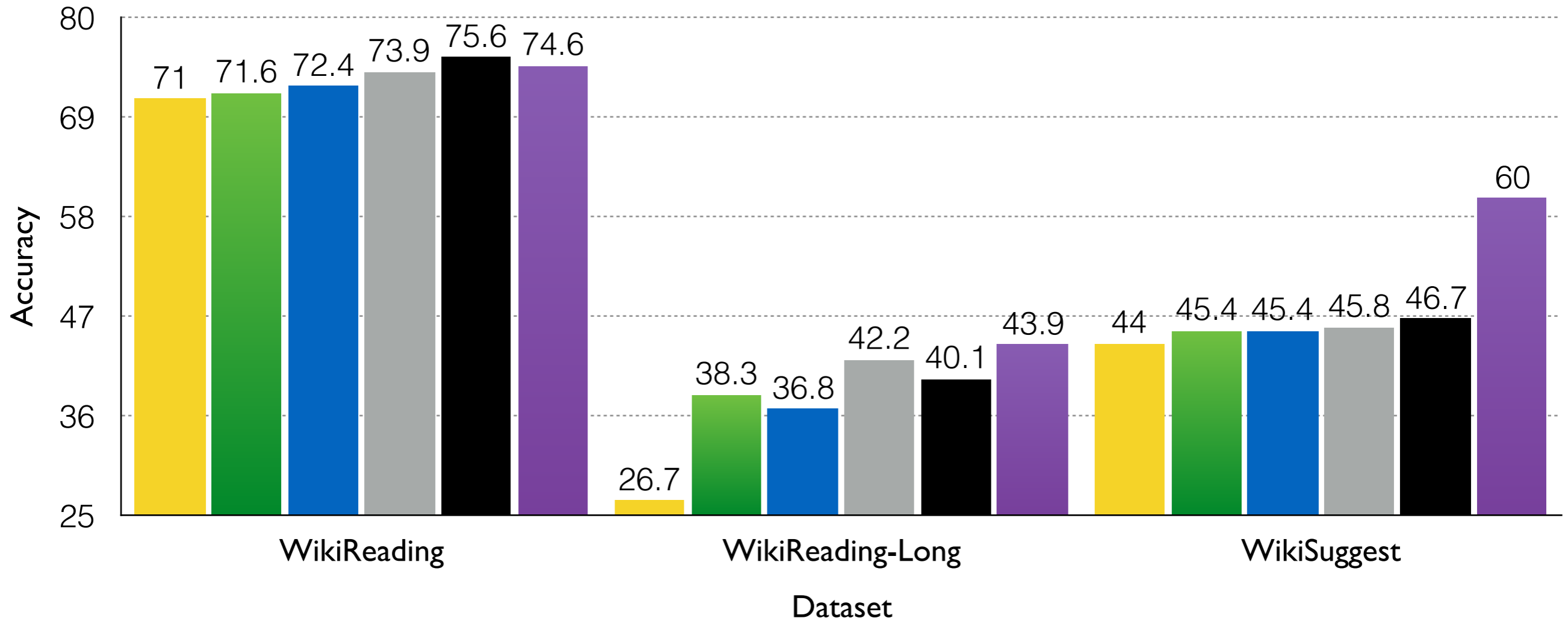
Comparison Systems

- 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.
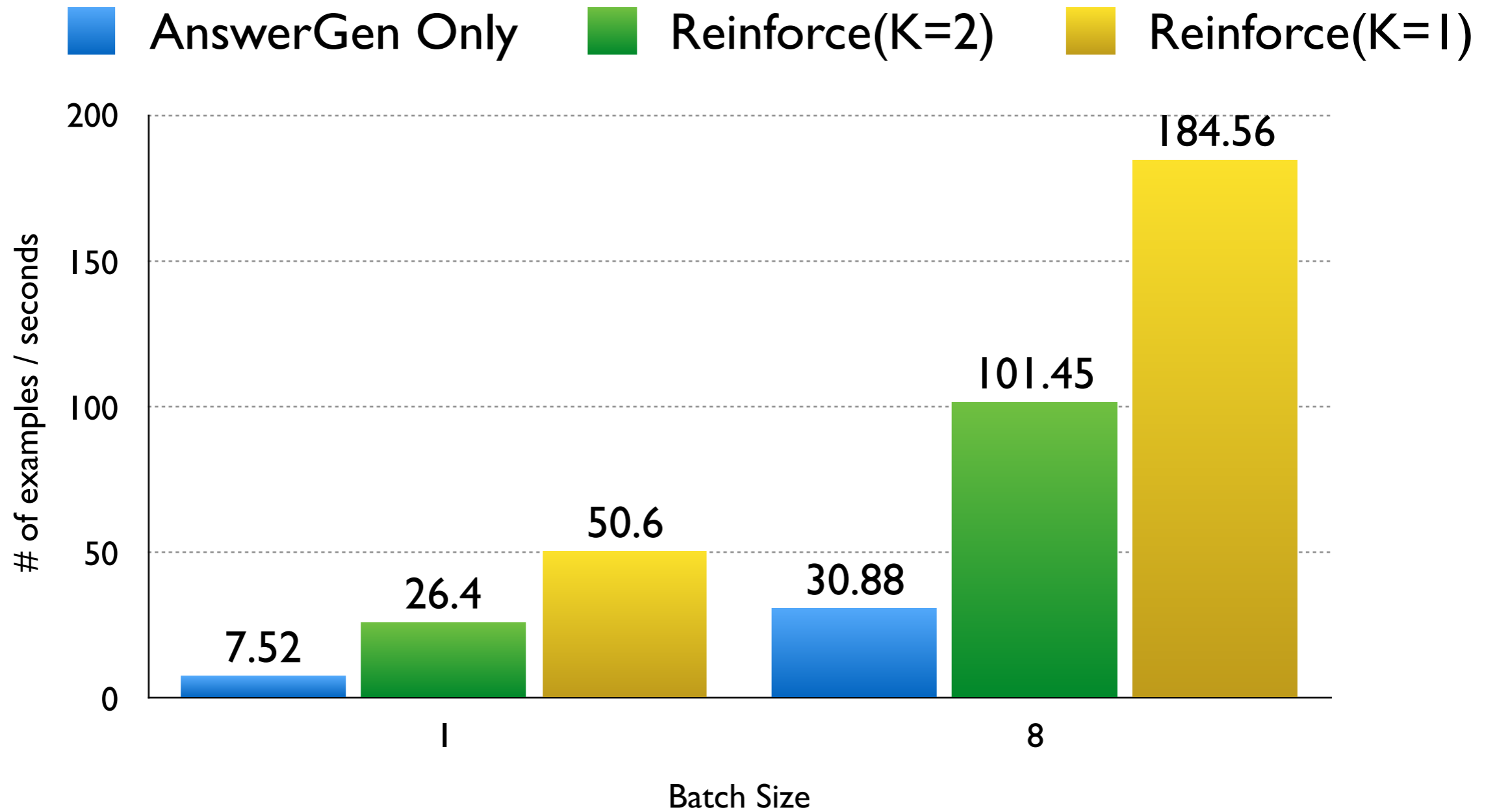
Accuracy Results

Accuracy Results

Legend: First (yellow), SoftAttend (green), Pipeline (blue), Reinforce (gray), AnswerGen Only (black), Heuristic Oracle (purple)

WikiReading: First 71, SoftAttend 71.6, Pipeline 72.4, Reinforce 73.9, AnswerGen Only 75.6, Heuristic Oracle 74.6

WikiReading-Long: First 26.7, SoftAttend 38.3, Pipeline 36.8, Reinforce 42.2, AnswerGen Only 40.1, Heuristic Oracle 43.9

WikiSuggest: First 44, SoftAttend 45.4, Pipeline 45.4, Reinforce 45.8, AnswerGen Only 46.7, Heuristic Oracle 60

Accuracy / Dataset

Speed Results

Speed Results
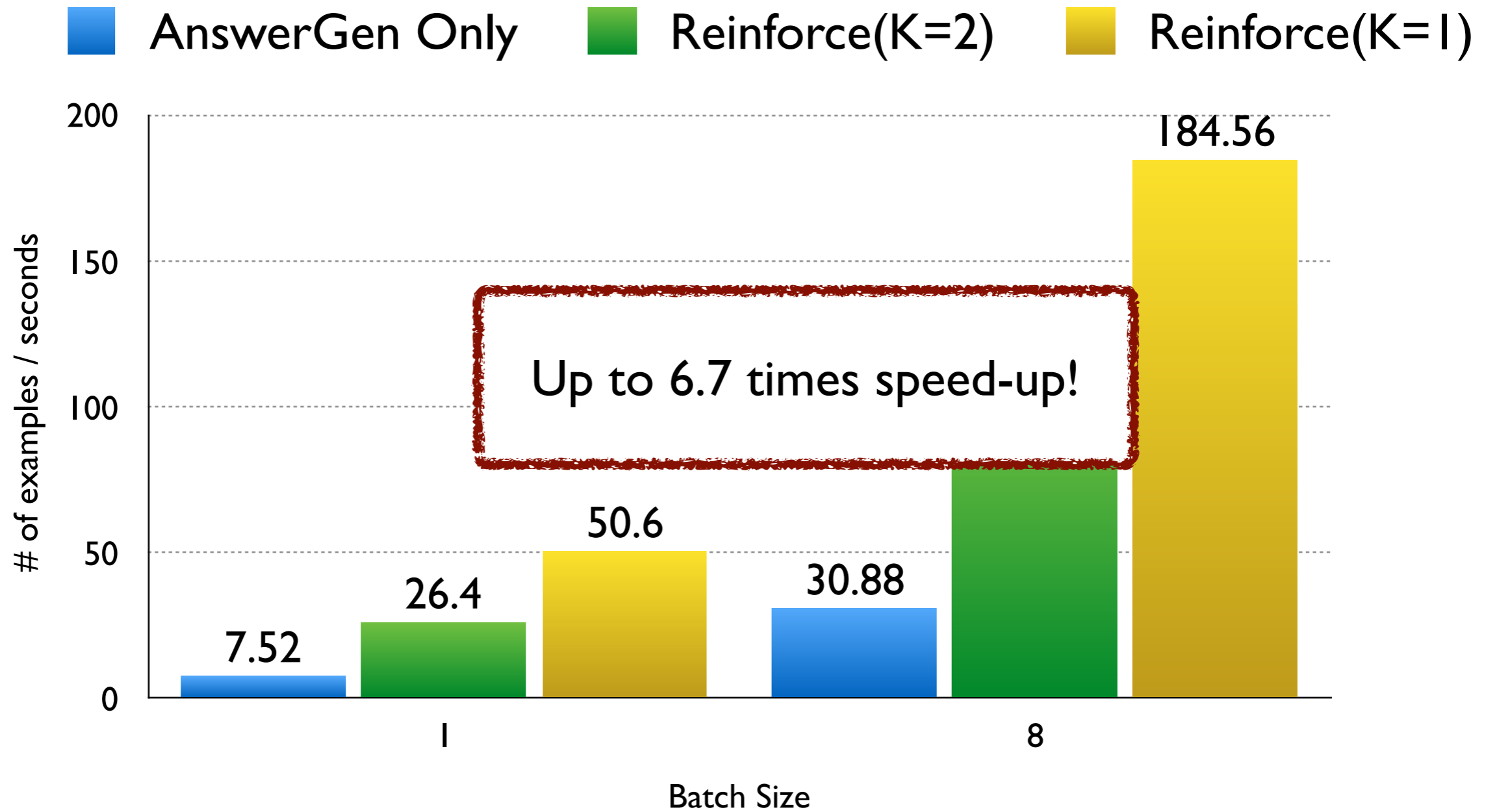
Error Analysis

	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

Future Work

- Introduce more flexible document selection

- Extend to more challenging dataset (multi document setting)

Thank you!
Questions?

Sentence Selection Accuracy