Denoising Distantly Supervised Open-Domain Question Answering

Yankai Lin

THUNLP, Tsinghua University

Reading Comprehension

Reading comprehension is the ability to **read**, **process** and **understand** natural language text.

Reading Comprehension

Question + Passage ⟶ Answer

Question: What are injectors used to supply?

Passage: The Rankine cycle and most practical steam engines have a water pump to recycle or top up the boiler water, so that they may be run continuously. Utility and industrial boilers commonly use multi-stage centrifugal pumps; however, other types are used. Another means of supplying lower-pressure boiler feed water is an injector, which uses a steam jet usually supplied from the boiler. Injectors became popular in the 1850s but are no longer widely used, except in applications such as steam locomotives.

Answer: lower-pressure boiler feed water

Reading Comprehension

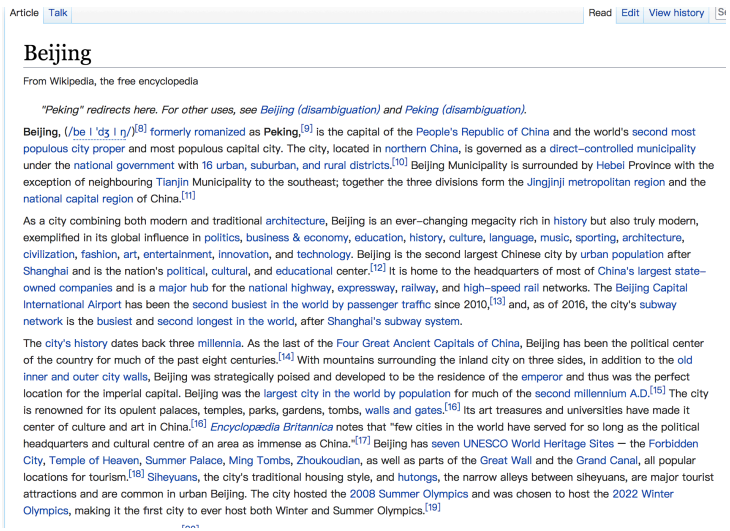
- People have proposed massive reading comprehension models and achieved promising results
  - BiDAF (Seo et al. 2016)
  - Attentive Reader (Chen et al., 2016)
  - AoA Reader (Cui et al., 2017)
  - R-NET (Wang et al., 2017)
  - …
- Problem
  - Rely on pre-identified relevant text, not practical in real-world

Open Domain Question Answering (OpenQA)

Question → Passage → Answer

Question: What's the population in Beijing?

Passage:



Article Talk | Read Edit View history

Beijing

From Wikipedia, the free encyclopedia

"Peking" redirects here. For other uses, see Beijing (disambiguation) and Peking (disambiguation).

Beijing, (/beɪ ˈdʒɪ ŋ/)[8] formerly romanized as **Peking**,[9] is the capital of the People's Republic of China and the world's second most populous city proper and most populous capital city. The city, located in northern China, is governed as a direct-controlled municipality under the national government with 16 urban, suburban, and rural districts.[10] Beijing Municipality is surrounded by Hebei Province with the exception of neighbouring Tianjin Municipality to the southeast; together the three divisions form the Jingjinji metropolitan region and the national capital region of China.[11]

As a city combining both modern and traditional architecture, Beijing is an ever-changing megacity rich in history but also truly modern, exemplified in its global influence in politics, business & economy, education, history, culture, language, music, sporting, architecture, civilization, fashion, art, entertainment, innovation, and technology. Beijing is the second largest Chinese city by urban population after Shanghai and is the nation's political, cultural, and educational center.[12] It is home to the headquarters of most of China's largest state-owned companies and is a major hub for the national highway, expressway, railway, and high-speed rail networks. The Beijing Capital International Airport has been the second busiest in the world by passenger traffic since 2010,[13] and, as of 2016, the city's subway network is the busiest and second longest in the world, after Shanghai's subway system.

The city's history dates back three millennia. As the last of the Four Great Ancient Capitals of China, Beijing has been the political center of the country for much of the past eight centuries.[14] With mountains surrounding the inland city on three sides, in addition to the old inner and outer city walls, Beijing was strategically poised and developed to be the residence of the emperor and thus was the perfect location for the imperial capital. Beijing was the largest city in the world by population for much of the second millennium A.D.[15] The city is renowned for its opulent palaces, temples, parks, gardens, tombs, walls and gates.[16] Its art treasures and universities have made it center of culture and art in China.[16] *Encyclopædia Britannica* notes that "few cities in the world have served for so long as the political headquarters and cultural centre of an area as immense as China."[17] Beijing has seven UNESCO World Heritage Sites — the Forbidden City, Temple of Heaven, Summer Palace, Ming Tombs, Zhoukoudian, as well as parts of the Great Wall and the Grand Canal, all popular locations for tourism.[18] Siheyuans, the city's traditional housing style, and hutongs, the narrow alleys between siheyuans, are major tourist attractions and are common in urban Beijing. The city hosted the 2008 Summer Olympics and was chosen to host the 2022 Winter Olympics, making it the first city to ever host both Winter and Summer Olympics.[19]

Search

Reading Comprehension

Answer: 21.148 million

Difference with Reading Comprehension

- Reading Comprehension
  - Input
    - Question & Passage
  - Passage
    - Pre-identified
    - Only one
    - Related to the question
- OpenQA
  - Input
    - Only question
  - Passage
    - Search results
    - Multiple
    - May not be related to the question

Open Domain Question Answering

- Researchers have made some attempts to answer open-domain questions.
  - DrQA (Chen et al. 2017)
  - R^3 (Wang et al. 2018)
  - …

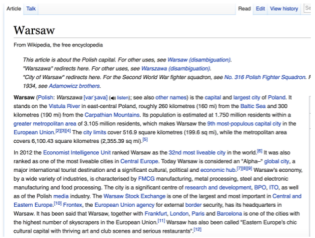
DrQA

- Document retriever + Reading comprehension

Open-domain QA
SQuAD, TREC, WebQuestions, WikiMovies

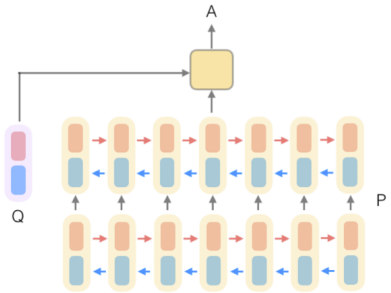
Q: How many of Warsaw's inhabitants spoke Polish in 1933?

Document Retriever



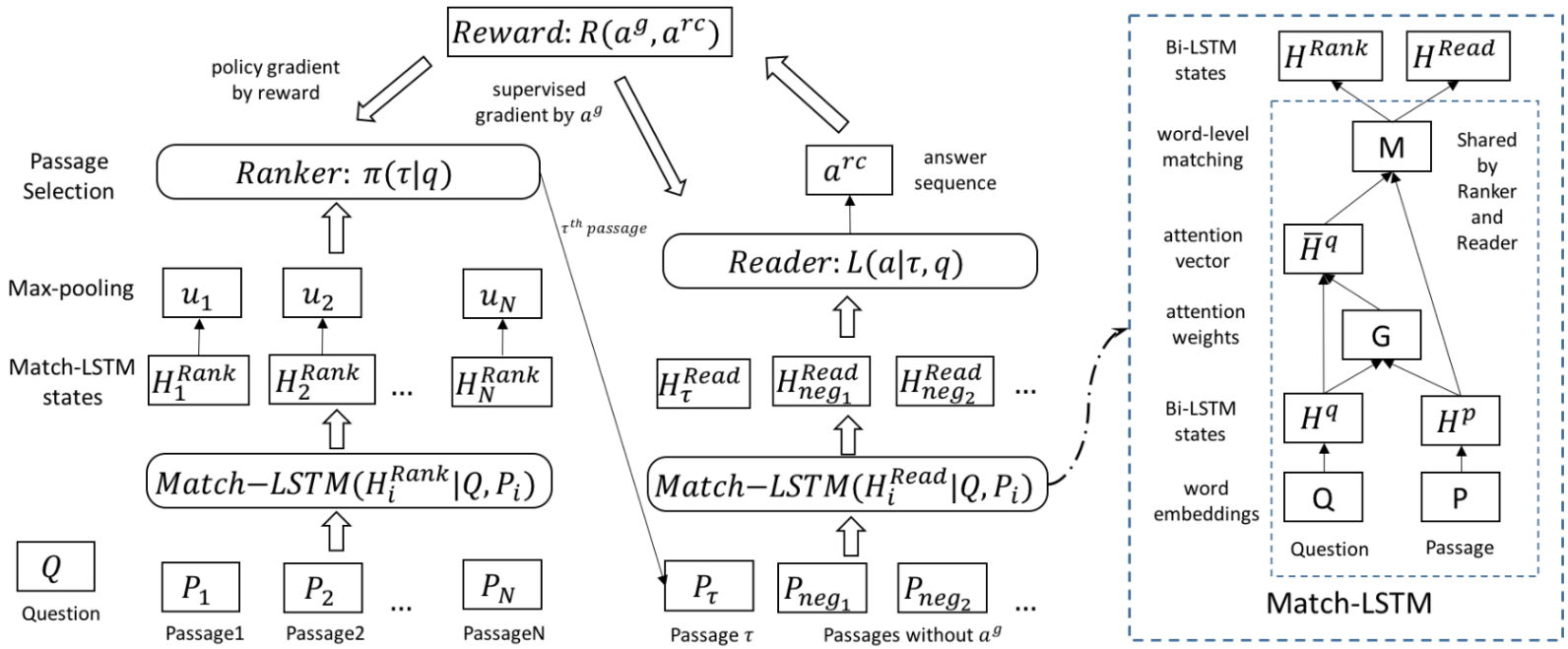
Document Reader

833,500



R^3

- Reinforcement learning



Open Domain Question Answering

- Researchers have made some attempts to answer open-domain questions.
  - DrQA (Chen et al. 2017)
  - R^3 (Wang et al. 2018)
  - …
- Problem
  - Cannot deal with noise problem of the retrieved passages
  - Cannot effectively aggregate information from different passages

Noise Problem in OpenQA

- Question
  - Which country has the fourth largest population?
- Passage
  - With well over 210 million people, Indonesia is the fourth most populous country in the world.
  - …, Indonesia is New Zealand's fourth largest source of imports.

- Not all retrieved passages are related to the question!

Aggregate Information in OpenQA

- Question
  - What famous artist could write with both his left and right hand at the same time?
- Passage
  - Leonardo Da Vinci was and is best known as an artist, …
  - … the reason Leonardo da Vinci used his left hand exclusively was that his right hand was paralyzed.
  - … forced me to use my right-hand, … beat my left-hand fingers with … so that I use the right hand.
- Need to aggregate information from all paragraphs!

Motivation

How human being read?

Fast Skimming + Careful Reading + Summarizing

- Fast skimming aims to identify relevant text from large-scale corpus.
- Careful reading aims to extract answers from a specified relevant text.
- Summarizing aims to aggregate information of all relevant text.

Our Model

Question:

What's the capital of Ireland?

Search

Paragraphs:

p_1: As the capital of Ireland, Dublin is …
p_2: Ireland is an island in the North Atlantic…
p_3: Dublin is the capital of Ireland. Besides, Ottawa is one of famous tourist cities in Ireland and ...

Summarizing $P(a|q,P)=\sum_{p_i \in P} P(a|q,p_i)P(p_i|q,P)$

Fast Skimming

p_1: As the capital of Ireland, Dublin is …
p_3: Dublin is the capital of Ireland. Besides, Ottawa is one of famous tourist cities in Ireland and ...

Paragraph Selector

$P(p_i|q,P)$

Careful Reading

p_1: As the capital of Ireland, Dublin is …
p_3: Dublin is the capital of Ireland. Besides, Dublin is one of famous tourist cities in Ireland and ...

Paragraph Reader

$P(a|q,p_i)$

Answer: Dublin

Paragraph & Question Encoding

- Word Representation
  - Word embedding (pretrained by GloVe)
  - Aligned question embedding (only for paragraphs)
  - Exact match (only for paragraphs)
- Unified Encoder
  - MLP $\hat{\mathbf{q}}_i^j = \mathbf{MLP}(\mathbf{q}_i^j)$
  - RNN $\{\hat{\mathbf{q}}^1, \hat{\mathbf{q}}^2, \cdots, \hat{\mathbf{q}}^{|q|}\} = \mathbf{RNN}(\{\mathbf{q}^1, \mathbf{q}^2, \cdots, \mathbf{q}^{|q|}\})$
- Question Encoder

$$\hat{\mathbf{q}} = \sum_j \alpha^j \hat{\mathbf{q}}^j, \quad \alpha_i = \frac{\exp(\mathbf{w}_b \mathbf{q}_i)}{\sum_j \exp(\mathbf{w}\mathbf{q}_j)}$$

Paragraph Selector

- Motivation: to filter out noisy paragraphs to aggregate useful information
- Measure the probability of each paragraph containing the answer
- A max layer and a softmax layer

$$\Pr(p_i|q, P) = \text{softmax}\left(\max_j(\hat{\mathbf{p}}_i^j \mathbf{W}\mathbf{q})\right)$$

Paragraph Reader

- Extract answer of the question from a given paragraph
- Calculate the start and end position of the answer span

$$P_s(j) = \text{softmax}(\bar{\mathbf{p}}_i^j \mathbf{W}_s \bar{\mathbf{q}}), \qquad P_e(j) = \text{softmax}(\bar{\mathbf{p}}_i^j \mathbf{W}_e \bar{\mathbf{q}})$$

- Multiple answer span problem in OpenQA
  - Max
    - Only one answer span indicates the answer

$$\Pr(a|q, p_i) = \max_j \Pr_s(a_s^j) \Pr_e(a_e^j)$$

  - Sum
    - All answer spans is the same

$$\Pr(a|q, p_i) = \sum_j \Pr_s(a_s^j) \Pr_e(a_e^j)$$

Experimental Setup

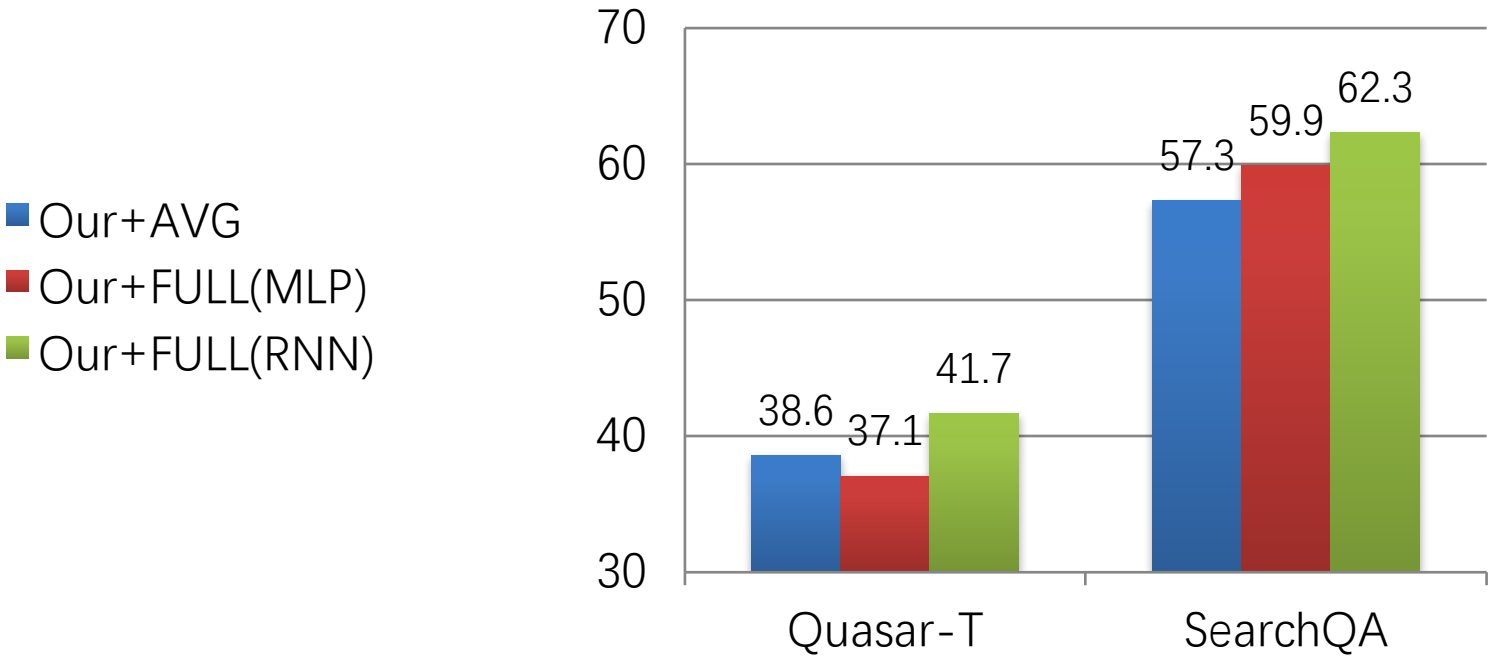
- Data
  - Quasar-T (Dhingra et al., 2017)
  - SearchQA (Dunn et al., 2017)
  - TriviaQA (Joshi et al., 2017)

Datasets	#Train	#Dev	#Test
Quasar-T	28,496	3,000	3,000
SearchQA	99,811	13,893	27,247
TriviaQA	66,828	11,313	10,832

- Evaluation
  - EM, F1

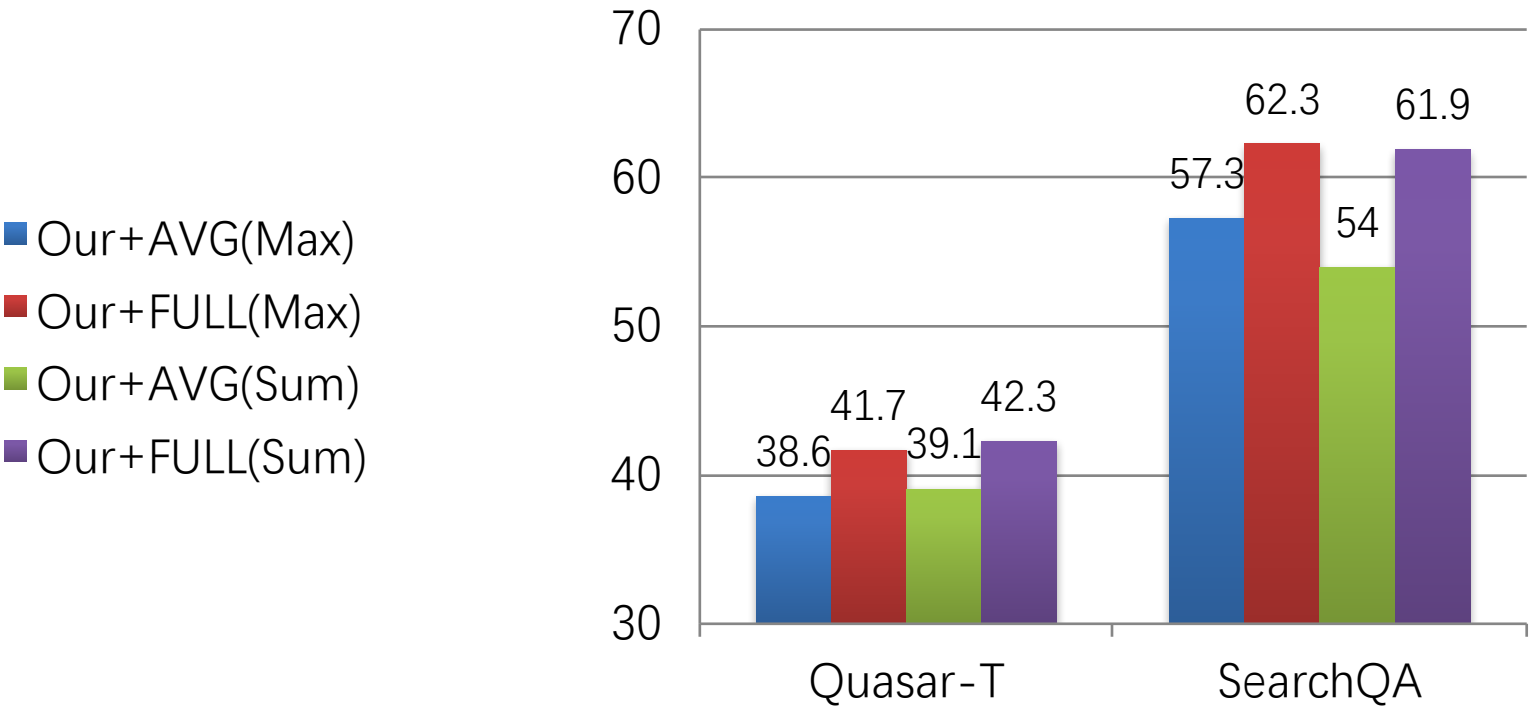
Effect of Different Paragraph Selectors

- RNN selector is better!



Effect of Different Paragraph Readers

- Max and Sum reader is comparable.



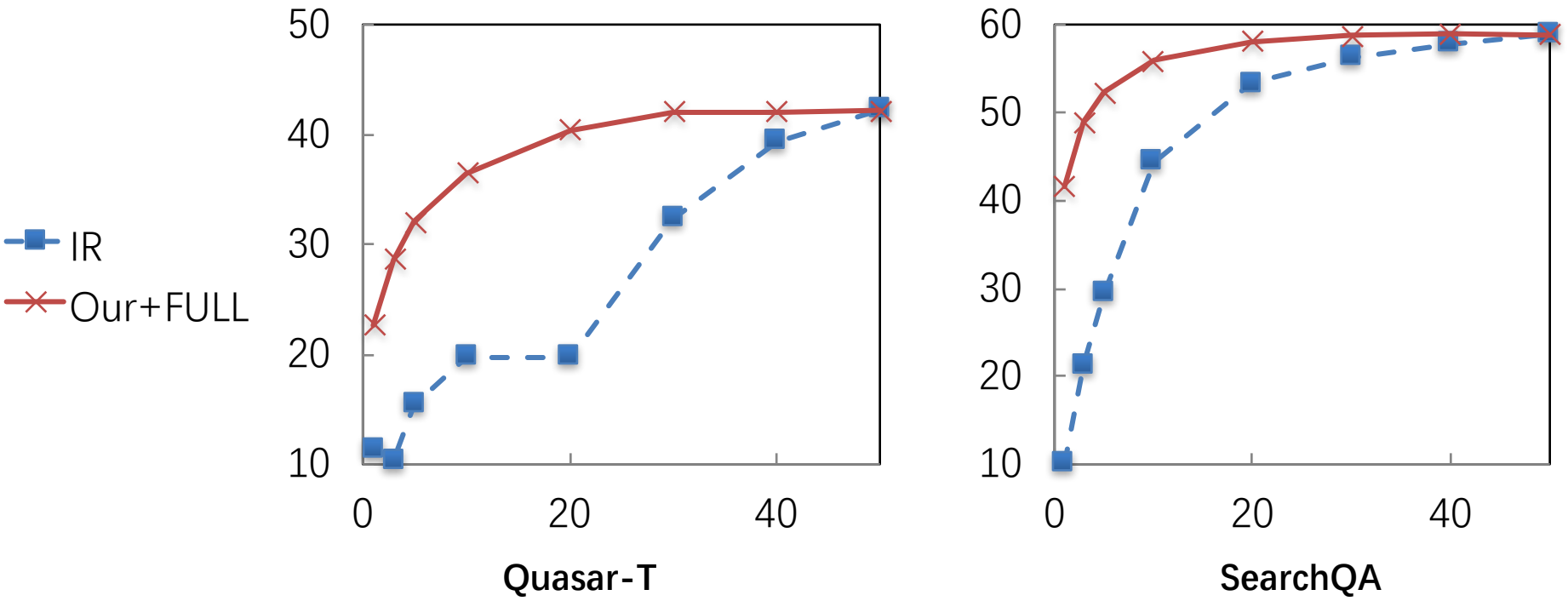
Overall Results

- Quasar-T: +8 points
- SearchQA: +9 points
- TriviaQA: +3 points

Datasets	Quasar-T		SearchQA		TriviaQA	
Models	EM	F1	EM	F1	EM	F1
GA (Dhingra et al., 2017a)	26.4	26.4	-	-	-	-
BiDAF (Seo et al., 2017)	25.9	28.5	28.6	34.6	-	-
AQA (Buck et al., 2017)	-	-	40.5	47.4	-	-
R^3 (Wang et al., 2018a)	35.3	41.7	49.0	55.3	47.3	53.7
Our + AVG	38.5	45.7	55.6	61.0	42.7	48.2
+ FULL	**42.2**	**49.3**	**58.8**	**64.5**	**48.7**	**56.3**

Performance with different numbers of paragraphs

- Our model performs better with a few paragraphs!



Potential improvement

- Our model is more potential using answer re-ranking.

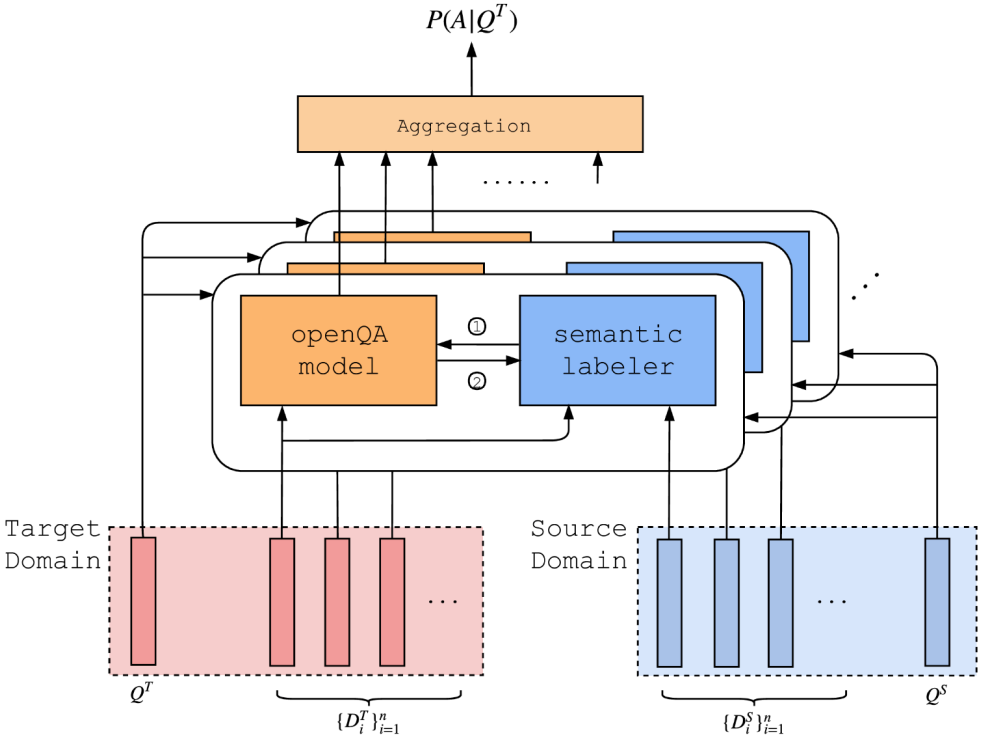
Datasets		Quasar-T		SearchQA	
Models	Top-k	EM	F1	EM	F1
R^3	1	35.5	41.6	51.2	57.3
	3	46.2	53.5	63.9	68.9
	5	51.0	58.9	69.1	73.9
Our+FULL	1	42.2	49.3	58.8	67.4
	3	53.1	62.0	72.9	77.4
	5	56.4	66.4	76.9	81.0

Motivation

- Distantly supervised OpenQA datasets lacks enough supervised signal to learn good paragraph selector (discriminator)
- Can we leverages alignment information from query-sentence pairs in supervised reading comprehension (RC) datasets for enhance?

Our Model

- Regard it as a transfer learning problem
  - OpenQA model
  - Semantic labeler



Semantic Labeler

- Transfer knowledge from the supervised RC dataset
- Two Strategies
  - Semi-supervised Learning with Semantic Labels (SSL)
    - Train a semantic labeler with RC dataset

$$\mathcal{L}_{SL} = \frac{1}{n'} \sum_{i=1}^{n'} -y_i^{\mathcal{S}} \log(\hat{y}_i^{\mathcal{S}}) - (1 - y_i^{\mathcal{S}}) \log(1 - \hat{y}_i^{\mathcal{S}}).$$

    - Use the semantic labeler to give soft label for OpenQA dataset

$$\mathcal{L}_{WD} = \frac{1}{n} \sum_{i=1}^{n} -\hat{y}_i^{\mathcal{T}} \log(R_i^{\mathcal{T}}) - (1 - \hat{y}_i^{\mathcal{T}}) \log(1 - R_i^{\mathcal{T}})$$

  - Collaborative Learning with Semantic Labels (CSL)
    - Collaborative learning for semantic labeler and paragraph selector

$$\mathcal{L}_{TL} = \frac{1}{n} \sum_{i=1}^{n} -R_i \log(\hat{y}_i^{\mathcal{T}}) - (1 - R_i) \log(1 - \hat{y}_i^{\mathcal{T}})$$

Overall Results

- Achieve state-of-the-art performance in all datasets

Datasets	Quasar-T		SearchQA		TriviaQA	
Models	EM	F1	EM	F1	EM	F1
Denoise OpenQA (Lin et al., 2018)	42.2	49.3	58.8	64.5	48.7	56.3
Re-ranker (Wang et al., 2018)	42.3	49.6	57.0	63.2	50.6	57.3
S-Norm (Clark & Gardner, 2018)					61.6	67.6
SSL	61.4	66.6	**59.5**	**65.1**	61.9	66.4
CSL	**62.2**	**67.5**	59.4	64.9	**63.7**	**68.2**

Performance of Sentence Discriminator

- Great improvement to measure if a paragraph contains the answer span by incorporating the information from supervised RC datasets

Datasets	**TriviaQA(unfiltered)**	**Quasar-T**	**SearchQA**
Models	Top1	Top1	Top1
Paragraph Selector (Lin et al., 2018)	-	27.7	58.9
Semantic Labeler (pretrained on SQuAD)	38.8	34.7	52.3
Sentence Discriminator + DISTANT	54.4	59.3	71.6
Sentence Discriminator + SEMANTIC	**57.4**	**62.6**	**72.6**

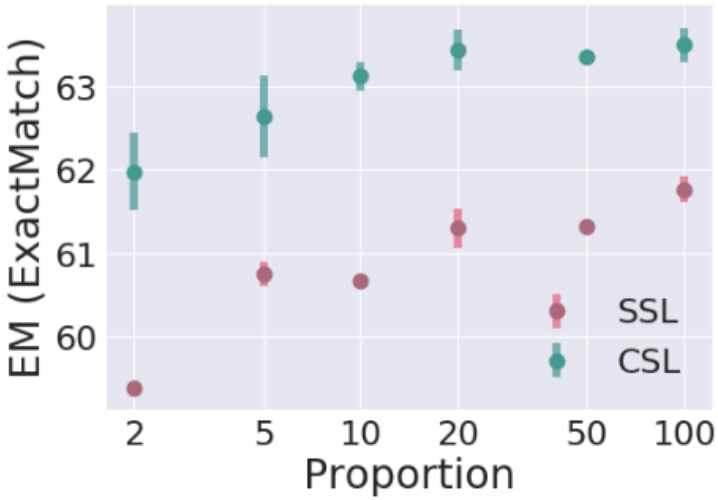
Case Study

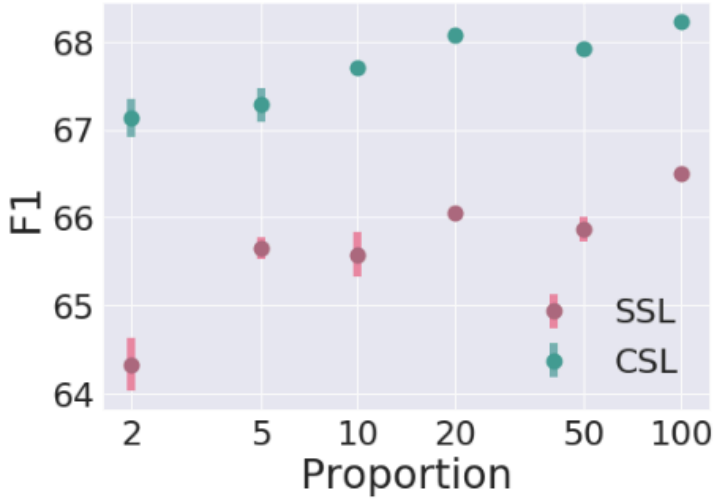
- Semantic labels could give a better estimation compared with distant supervision labels.

question: *Which sport has a name which literally means 'gentle way'?*	**label**	
Ground truth: judo	**distant**	**semantic**
The term "do way", which is used in the names of arts like **judo**, aikido ...	1	0.53
Sport and beyond despite the literal meaning of **judo** being 'gentle way' ...	1	0.94
Kano took the name judo from jikishin ryu **judo**, which is an older school ...	1	0.34
Dr. Kano meant for his gentle way to be a way to live, a path to follow.	0	0.91

Different proportion of supervised data

- CSL model is more robust with few supervised data





Conclusion

- We model how human being’ s read
  - Fast skimming + Careful reading + Summarizing
- We transfer supervised RC dataset to OpenQA model
- Our system has promising performance only using a few paragraphs
- Our system can be further improved by answer re-ranking

Future Direction

- Incorporate knowledge
- Transfer to other area
- Consider more complex reasoning
- Question rewriting
- Transfer to task-orient QA

Thank you!