Survey of Multiple Choice Question Answering Systems

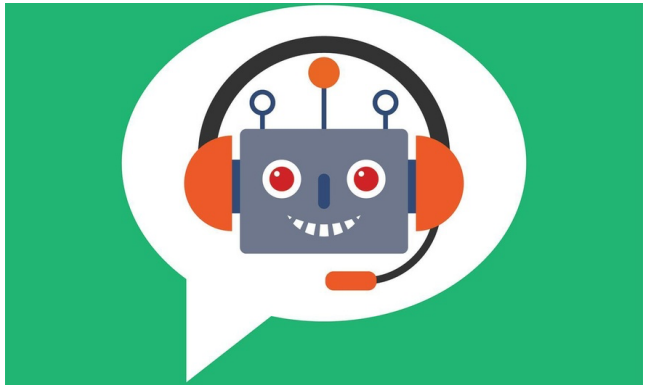
Vikas Yadav

THE UNIVERSITY OF ARIZONA

Computational Language Understanding

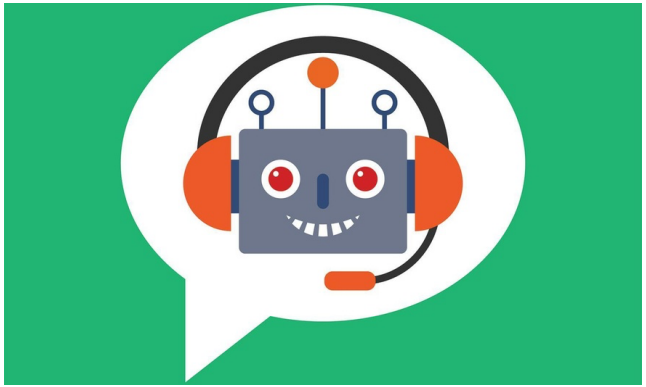
Task: question answering

Which instrument measures atmospheric pressure?



Task: question answering

Which instrument measures atmospheric pressure?



Barometer

Flavors of QA: Multiple-choice QA

AI2's Reasoning Challenge (ARC) (Clark et al., 2018)

Question: Which change in the state of water particles causes the particles to become arranged in a fixed position ?

A] Boiling B] Melting **C] Freezing** D] Evaporating

Flavors of QA: Multi-hop QA

Multiple sentence reading comprehension (MultiRC) (Kashabi et al., 2018)

(SENT 0): The driver of the Facel Vega car, Michel Gallimard, who was Camus's publisher and close friend, also died in the accident.", (SENT 1): "In August 2011, the Milan newspaper Corriere della Sera reported a theory that the writer had been the victim of a Soviet plot, but Camus's biographer, Olivier Todd, did not consider it credible.", (SENT 2): 'Camus was buried in the Lourmarin Cemetery, Lourmarin, Vaucluse, France.', (SENT 3): 'He was the second-youngest recipient, at the age of 44, of the Nobel Prize in Literature, after Rudyard Kipling, at the age of 42.', (SENT 4): 'He was survived by his wife and twin son and daughter, Jean and Catherine, who hold the copyrights to his work.', (SENT 5): "Two of Camus's works were published posthumously.", (SENT 6): "The first, entitled A Happy Death (1970), featured a character named Patrice Meursault, comparable to The Stranger's Meursault.", (SENT 7): 'There is scholarly debate as to the relationship between the two books.', (SENT 8): **'The second was an unfinished novel, The First Man (1995), which Camus was writing before he died**.', (SENT 9): **'The novel was an autobiographical work about his childhood in Algeria**.

Question - Which novel did Camus write about his childhood in Nigeria?

Answer - The First Man

Flavors of QA: sentence selection

WikiQA (Yang et al., 2015)

Question: how does a water pump work?

A large, electrically driven pump (electropump) for waterworks near the Hengsteysee , Germany.

Pumps operate by some mechanism (typically reciprocating or rotary), and consume energy to perform mechanical work by moving the fluid.

A pump is a device that moves fluids (liquids or gases), or sometimes slurries , by mechanical action.

Flavors of QA: reading comprehension

SQuAD (Rajpurkar et al., 2016)

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity. The main forms of precipitation include drizzle, rain, sleet, snow, graupel and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals **within a cloud**. Short, intense periods of rain in scattered locations are called “showers”.

Question - Where do water droplets collide with ice crystals to form precipitation?

Answer – “within a cloud”

Flavors of QA: cloze style

Context: 1 So they had to fall a long way . 2 So they got their tails fast in their mouths . 3 So they could n't get them out again . 4 That 's all . 5 `` Thank you , " said Alice , `` it 's very interesting . 6 I never knew so much about a whiting before . " 7 `` I can tell you more than that , if you like , " said the Gryphon . 8 `` Do you know why it 's called a whiting ? " 9 `` I never thought about it , " said Alice . 10 `` Why ? " 11 `` IT DOES THE BOOTS AND SHOES . ' 12 the Gryphon replied very solemnly . 13 Alice was thoroughly puzzled . 14 `` Does the boots and shoes ! " 15 she repeated in a wondering tone . 16 `` Why , what are YOUR shoes done with ? " 17 said the Gryphon . 18 `` I mean , what makes them so shiny ? " 19 Alice looked down at them , and considered a little before she gave her answer . 20 `` They 're done with blacking , I believe .

" **Query**: `` Boots and shoes under the sea , " the _____ went on in a deep voice , `` are done with a whiting ".

Candidates: Alice|BOOTS|Gryphon|SHOES|answer|fall|mouths|tone|way|whiting

Answer: gryphon

Flavors of QA: knowledge based QA

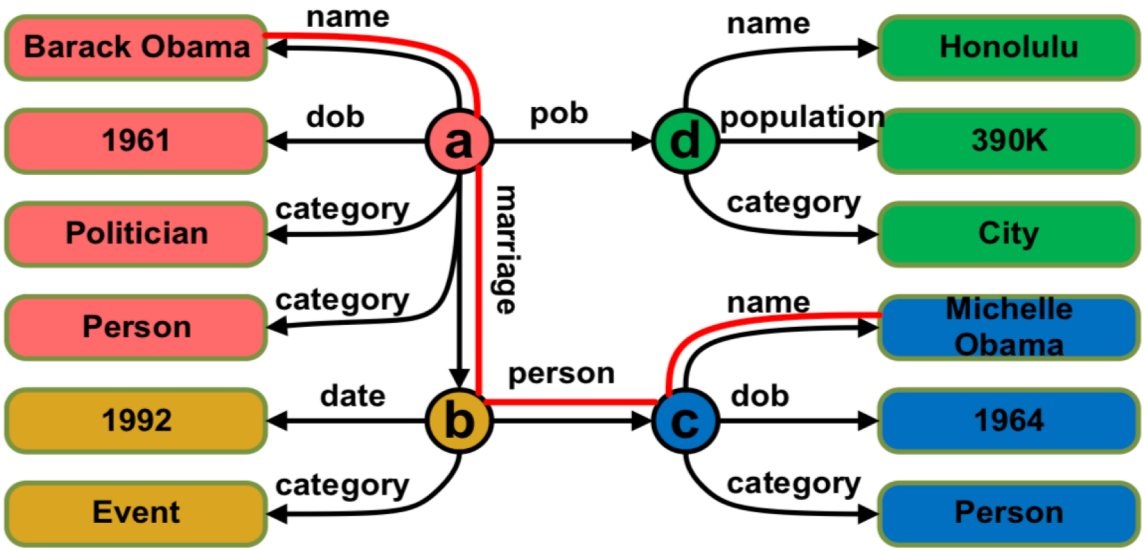


Figure 1: A toy RDF knowledge base (here, "dob" and "pob" stand for "date of birth" and "place of birth" respectively). Note that the "spouse of" intent is represented by multiple edges: name - marriage - person - name.

Image taken from Cui et al. (2017)

Flavors of QA: Maths QA

Closed-vocabulary algebra: Suppose $3x + y = 15$, where x is a positive integer. What is the difference between the largest possible value of y and the smallest possible value of x, assuming that y is also a positive integer?

Open-vocabulary algebra: At a basketball tournament involving 8 teams, each team played 4 games with each of the other teams. How many games were played at this tournament?

Geometry: The lengths of two sides of a triangle are $(x - 2)$ and $(x + 2)$, where $x > 2$. Which of the following ranges includes all and only the possible values of the third side y?

(A)$0<y<x$(B)$0<y<2x$(C) $4 < y < 2x$

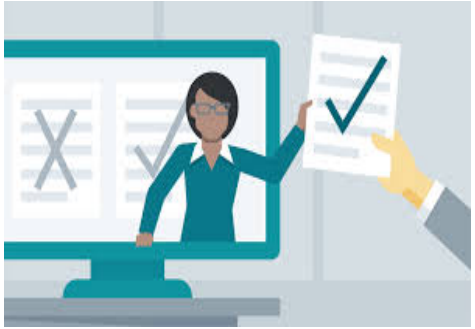
Example taken from Hopkins et al. (2019)

Our focus – Multiple choice QA

Information Systems

Data Sciences

Applications:



We have seen numerous systems

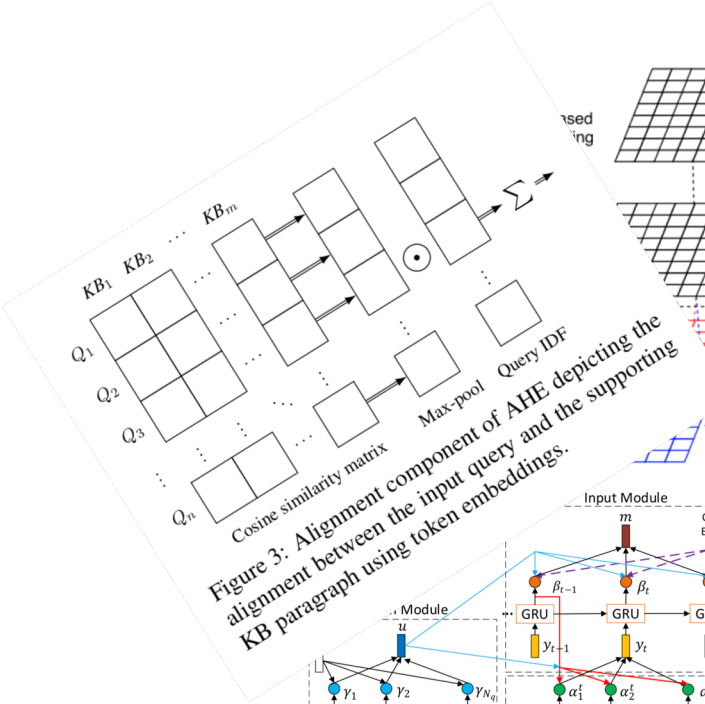
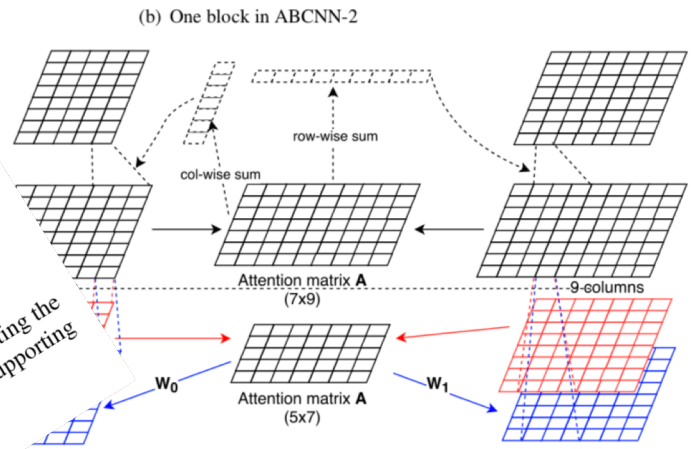
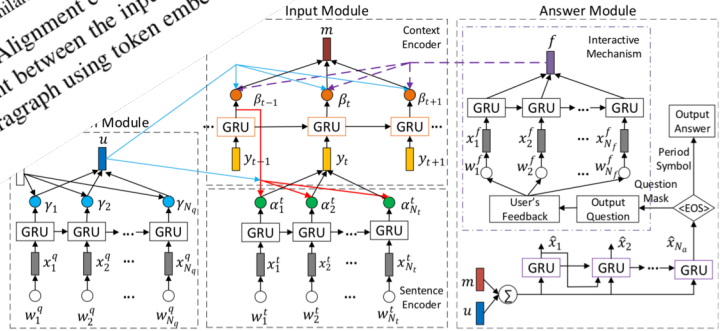
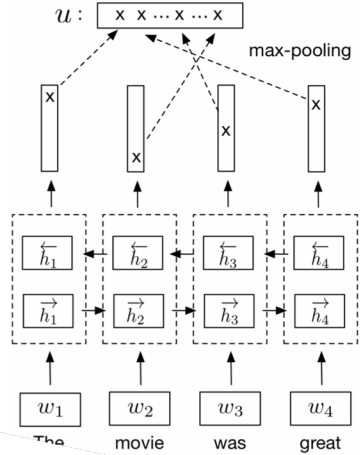
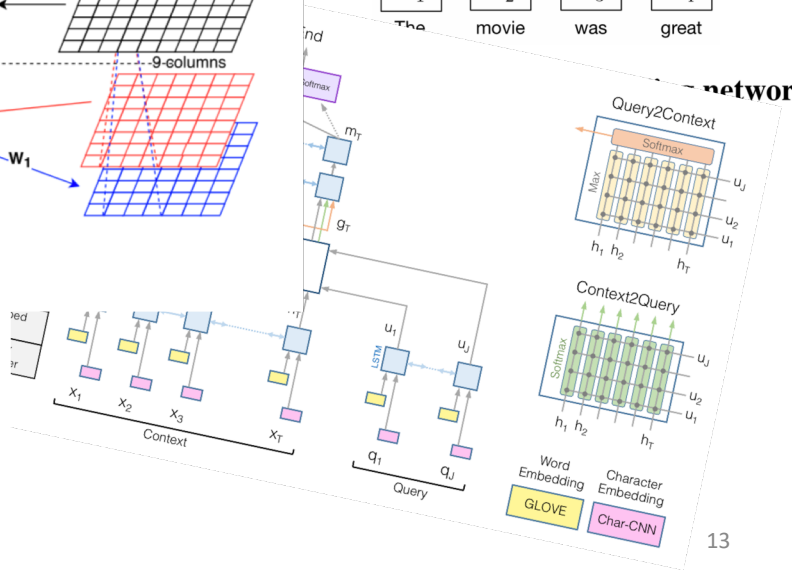


Figure 3: Alignment component of AHE depicting the alignment between the input query and the supporting KB paragraph using token embeddings.









But…

- Which one to select ?
- What are the pros and cons ?
- Is the MCQA training data unbiased and easily available ?

Hence…

- We propose a survey comparing various MCQA systems both quantitatively and qualitatively

- Highlight the anomalies in various MCQA datasets

MCQA datasets

ARC, 4th and 8th grade Science QA dataset

Question] George wants to warm his hands quickly by rubbing them. Which skin surface will produce the most heat?

1. Dry palms
2. Wet palms
3. Palms covered with oil
4. Palms covered with lotion

Multi-hop MCQA dataset (OpenBookQA, Mihaylov et.al, 2018)

Question] Which of these would let the most heat travel through?

1] New pair of jeans

2] A steel spoon in a cafeteria

3] A cotton candy at a store

4] A calvin klein cotton hat

- Science Fact*: Metal is a thermal conductor.*
- Common Knowledge: *Steel is made of metal.*
  Heat travels through a thermal conductor.

Anomalies in MCQA datasets

Train \ Test	Anamolies	Impact
ARC	There are many more supporting facts for the correct answers when compared to the incorrect	High ARC easy scores with simple IR methods (e.g., BM25) (Yadav et al., 2019a)
OpenBookQA	Contains artificial non-natural language questions	state-of-the-art performance can be achieved by just using question and candidate answer text (Mihaylov et al., 2018) alone.
MultiRC	None, other than slightly smaller in size compared to other multi-hop datasets like HotPotQA	NN models like BERT achieve similar performance as that of simple logistic regression model (Khashabi et al., 2018a)
HotPotQA	Majority (91%) of the sentences participating in multi-hop inference are the first sentences.	NN based justification selection models overfits this particular pattern
WikiQA	Questions come only from one source - WikiPedia on which, majority of the word embeddings are trained	Models like BERT which are specifically pretrained on WikiPedia text achieve substantially higher gains (Yadav et al., 2019a)
YahooQA!	Correct candidate answers have longer sentence length	Simple heuristic of selecting longest answer as correct achieves state-of-the-art-results.
MedHop QA	Majority of the most frequently co-occuring entities are the correct answers	Frequently co-occuring heuristics achieves state-of-the-art (Welbl et al., 2018)

Table 1: Anamolies in the datasets.

Types of MCQA systems

Methods

- Statistical NLP (SNLP)
- Information Retreival (IR)
- Artificial Intelligence (AI)

Statistical NLP (SNLP)

- Machine Learning
- Feedforward and Recurrent Neural Networks (RNN) systems
- Transformers based – BERT, Xlnet, RoBERTa

Statistical NLP (SNLP)

Machine Learning – Support vector machines(SVM), hidden markov model HMM, etc.

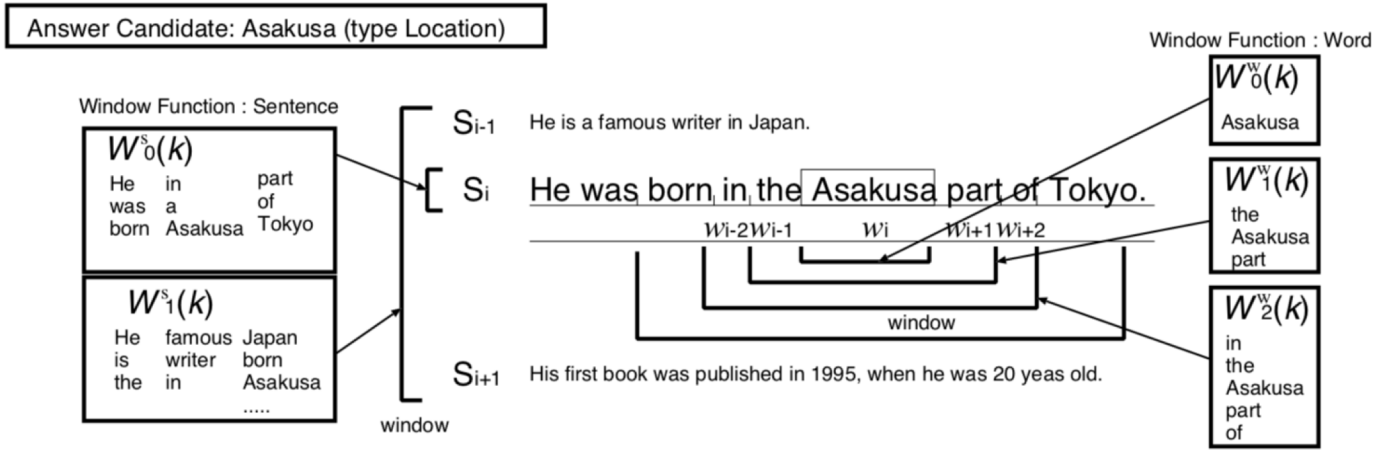


Figure 2: Example of Window Function

Image taken from Suzuki et al. (2002)

Statistical NLP (SNLP)

Feedforward and Recurrent Neural Networks Systems

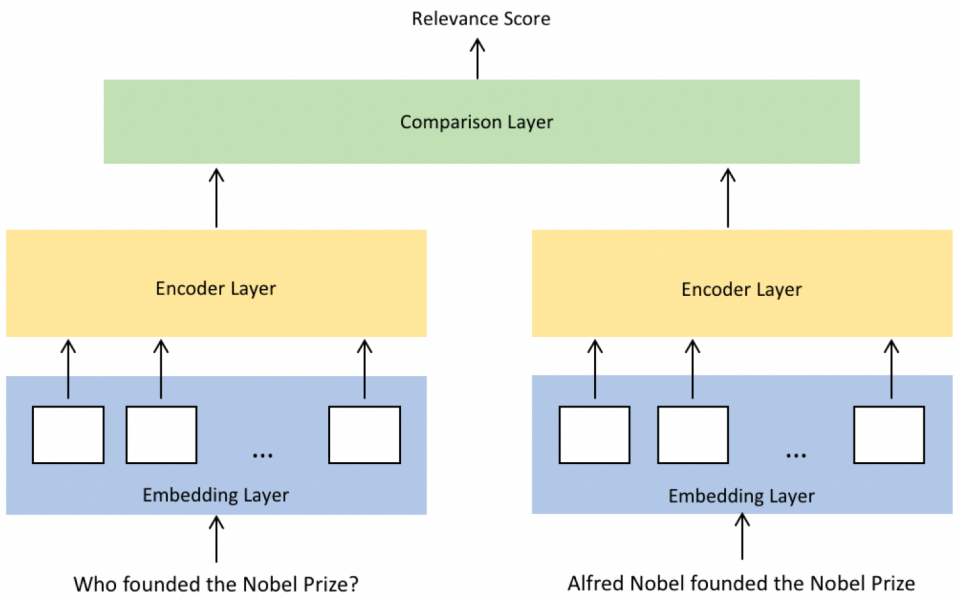


Image taken from Lai et al., (2018)

Statistical NLP (SNLP)

Transformers based – BERT, Xlnet, RoBERTa

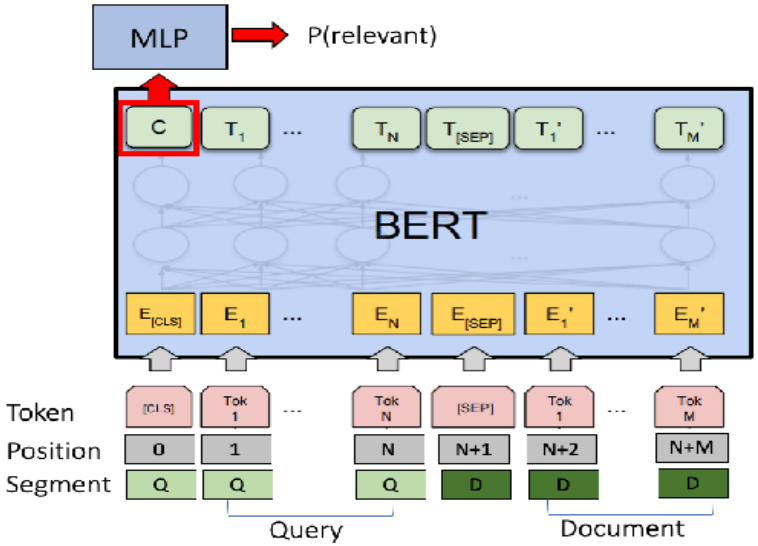


Image taken from link

Information Retrieval (IR)

- Conventional IR – BM25, PMI
- Alignment based IR – Sanity Check, AHE
- IR supporting SNLP - AutoROCC

Information Retrieval (IR)

Conventional IR – BM25, Point wise mutual information (PMI)

Information Retrieval (IR)

Alignment based IR – Sanity Check (SIGIR 18), AHE (NAACL 2019)

Question: What is a law enforcement murder book?

Correct Ans: In law enforcement parlance , the term murder book refers to the case file of a murder investigation.

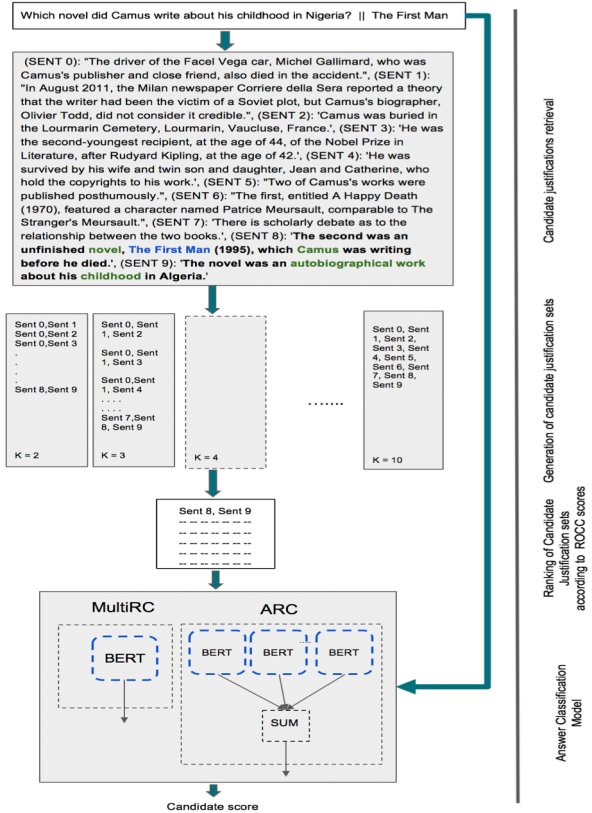
Incorrect Ans: Law enforcement agencies typically guard murder books carefully, and it is unusual for civilians to be given unfettered access to these kinds of records , especially for unsolved cases.

	Ranked Alignments	
Question Terms	Correct Answer Terms	Incorrect Answer Terms
law	[**law**, enforcement, …, parlance]	[**law**, enforcement, …, unfettered]
enforcement	[**enforcement**, law, …, book]	[**enforcement**, law, …, book]
murder	[**murder**, murder, …, file]	[**murder**, case, …, access]
book	[**book**, file, …, parlance]	[**book**, case, … unfettered]

Information Retrieval (IR)

IR supporting SNLP –

AutoROCC (EMNLP 2019)



Artificial Intelligence (AI)

- Integer Linear Programming
- Graph based QA models

Artificial Intelligence (AI)

Integer Linear Programming

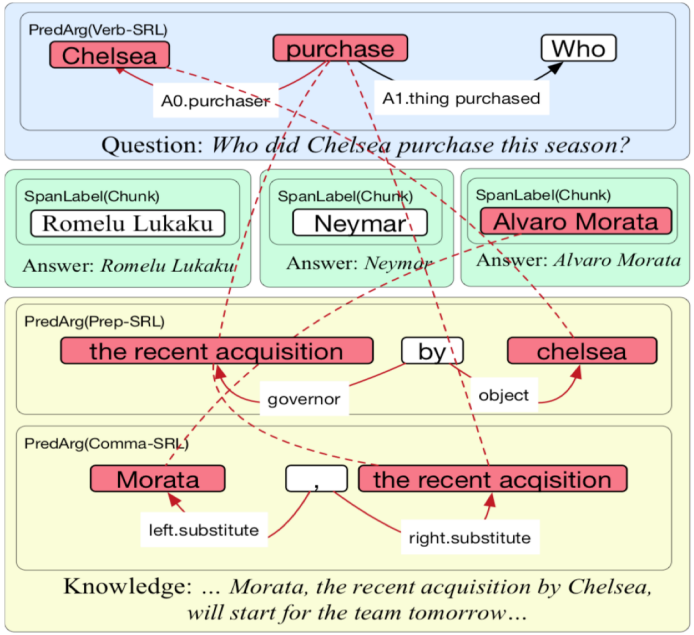


Image taken from Kashabi et al. (2018)

Artificial Intelligence (AI)

Graph based QA models

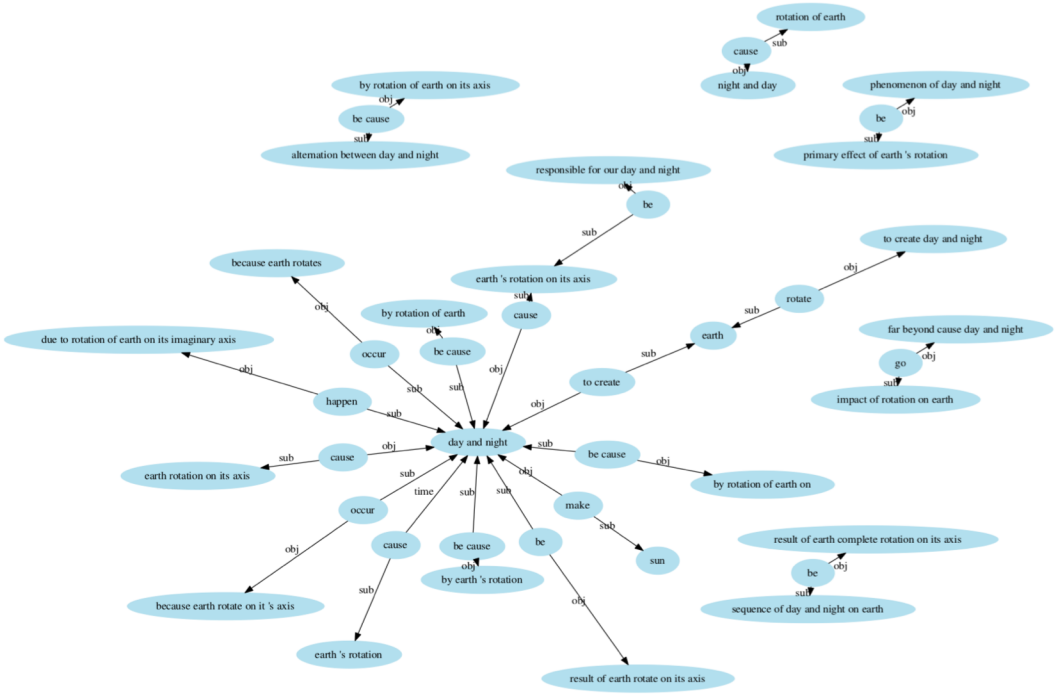


Image taken from Zhang et al. (2018)

Performance comparison

Performance comparison (ARC dataset)

#	Supervised ?	Type of KB	Model	Easy P@1	Challenge P@1
IR	No	Text	BM25 (IR Solver)	60	24
IR	No	Text	AHE (Yadav et al., 2019)	65.2	33.7
SNLP	Yes	--	Bi-LSTM max-out (Mihaylov et al., 2018)	34.3	33.9
SNLP	Yes	Text + Structured	ET-RR (Ni et al., 2018)	-	36.56
SNLP	Yes	Text	GPT - Reading Strategies (Sun et al., 2018)	63.5	33.9
SNLP	Yes	Text	AutoROCC+BERT (Yadav et al., 2019)	64.5	41.2
AI	Yes	Text + Structured	KG2 (Zhang et al., 2018)	-	31.7
AI	Yes	Text + Structured	DGEM (Clark et al., 2018)	59	27.1
AI	Yes	Text + Structured	Tuple-Inf (Clark et al., 2018)	60.7	23.8

Performance comparison

#	Supervised ?	Type of KB	Model	Easy P@1	Challen P@1
IR	No	Text	BM25 (IR Solver)	60	24
IR	No	Text	AHE (Yadav et al., 2019)	65.2	33.7
SNLP	Yes	--	Bi-LSTM max-out (Mihaylov et al., 2018)	34.3	33.9
SNLP	Yes	Text + Structured	ET-RR (Ni et al., 2018)	-	36.56
SNLP	Yes	Text	GPT - Reading Strategies (Sun et al., 2018)	63.2	42.2
SNLP	Yes	Text	AutoROCC+BERT (Yadav et al., 2019)	64.5	41.2
AI	Yes	Text + Structured	KG2 (Zhang et al., 2018)	-	31.7
AI	Yes	Text + Structured	DGEM (Clark et al., 2018)	59	27.1
AI	Yes	Text + Structured	Tuple-Inf (Clark et al., 2018)	60.7	23.8

Achieve good performance across both sets

Performance comparison

Fail terribly on Easy set

#	Supervised ?	Type of KB		Easy P@1	Challenge P@1
IR	No	Text		60	24
IR	No	Text	2019)	65.2	33.7
SNLP	Yes	--	Bi-LSTM max-out (Mihaylov et al., 2018)	34.3	33.9
SNLP	Yes	Text + Structured	ET-RR (Ni et al., 2018)	-	36.56
SNLP	Yes	Text	GPT - Reading Strategies (Sun et al., 2018)	63.2	42.2
SNLP	Yes	Text	AutoROCC+BERT (Yadav et al., 2019)	64.5	41.2
AI	Yes	Text + Structured	KG2 (Zhang et al., 2018)	-	31.7
AI	Yes	Text + Structured	DGEM (Clark et al., 2018)	59	27.1
AI	Yes	Text + Structured	Tuple-Inf (Clark et al., 2018)	60.7	23.8

Performance comparison

#	Supervised ?	Type of KB	Model	Easy P@1	Challenge P@1
IR	No	Text	BM25 (IR Solver)	60	24
IR	No	Text	AHE (Yadav et al., 2019)	65.2	33.7
SNLP	Yes	--	Bi-LSTM max-out (Mihaylov et al., 2018)	34.3	33.9
SNLP	Yes	Text + Structured	ET-RR (Ni et al., 2018)	-	36.56
SNLP	Yes	Text	GPT - Reading Strategies	63.2	42.2
SNLP	Yes	Text	AutoROCC+BERT (Yad	64.5	41.2
AI	Yes	Text + Structured	KG2 (Zhang et al	-	31.7
AI	Yes	Text + Structured	DGEM (Clark et al., 2018)	59	27.1
AI	Yes	Text + Structured	Tuple-Inf (Clark et al., 2018)	60.7	23.8

Fail terribly on Challenge set

Performance comparison

#	Supervised ?	Type of KB	Model	Easy P@1	Challenge P@1
IR	No	Te	5 (IR Solver)	60	24
IR	No	Te	adav et al., 2019)	65.2	33.7
SNLP	Yes	-	vlov et al., 2018)	34.3	33.9
SNLP	Yes	Text + Structured	ET-KR (Ni et al., 2018)	-	36.56
SNLP	Yes	Text	GPT - Reading Strategies (Sun et al., 2018)	63.2	42.2
SNLP	Yes	Text	AutoROCC+BERT (Yadav et al., 2019)	64.5	41.2
AI	Yes	Text + Structured	KG2 (Zhang et al., 2018)	-	31.7
AI	Yes	Text + Structured	DGEM (Clark et al., 2018)	59	27.1
AI	Yes	Text + Structured	Tuple-Inf (Clark et al., 2018)	60.7	23.8

Transformers Architecture indeed outperform previous approaches

Qualitative analysis – Domain Robustness

Domain robustness

Train\ Test	ARC Easy (P@1)	ARC Challenge (P@1)	WikiQA (MRR)
ARC Easy	34.3	23.1	38.7
ARC Challenge	27.0	**33.9**	39.1
WikiQA	25.8	24.3	67.4
AHE	64.6	**33.9**	68.5

Domain robustness of Bi-LSTM Maxout model (Mihaylov et al., 2018) which achieves state-of-the-art on ARC challenge.

Train\ Test	ARC Easy (P@1)	ARC Challenge (P@1)	WikiQA (MRR)
ARC Easy	38.8	24.1	53.9
ARC Challenge	36.2	26.4	41.5
WikiQA	38.4	25.4	**71.2**
AHE	64.6	33.9	68.5

Domain robustness of BiMPM model (Wang et al., 2017) which achieves one of the top performance on WikiQA.

Domain robustness

Train\ Test	ARC Easy (P@1)	ARC Challenge (P@1)	WikiQA (MRR)
ARC Easy	34.3	23.1	38.7
ARC Challenge	27.0	**33.9**	39.1
WikiQA	25.8	24.3	67.4
AHE	64.6	**33.9**	68.5

Domain transfer of Bi-LSTM Maxout (Mihaylov et al., 2018)

Train\ Test	ARC Easy (P@1)	ARC Challenge (P@1)	WikiQA (MRR)
ARC Easy	38.8	24.1	53.9
ARC Challenge	36.2	26.4	41.5
WikiQA	38.4	25.4	**71.2**
AHE	64.6	33.9	68.5

Domain transfer of BiMPM (Wang et al., 2017)

Domain robustness

Train\ Test	ARC Easy (P@1)	ARC Challenge (P@1)	WikiQA (MRR)
ARC Easy	34.3	23.1	38.7
ARC Challenge	27.0	**33.9**	39.1
WikiQA	25.8	24.3	67.4
AHE	64.6	**33.9**	68.5

Domain transfer of Bi-LSTM Maxout (Mihaylov et al., 2018)

Train\ Test	ARC Easy (P@1)	ARC Challenge (P@1)	WikiQA (MRR)
ARC Easy	38.8	24.1	53.9
ARC Challenge	36.2	26.4	41.5
WikiQA	38.4	25.4	**71.2**
AHE	64.6	33.9	68.5

Domain transfer of BiMPM (Wang et al., 2017)

Domain robustness

Train\ Test	ARC Easy (P@1)	ARC Challenge (P@1)	WikiQA (MRR)
ARC Easy	34.3	23.1	38.7
ARC Challenge	27.0	**33.9**	39.1
WikiQA	25.8	24.3	67.4
AHE	64.6	**33.9**	68.5

Domain transfer of Bi-LSTM Maxout (Mihaylov et al., 2018)

Train\ Test	ARC Easy (P@1)	ARC Challenge (P@1)	WikiQA (MRR)
ARC Easy	38.8	24.1	53.9
ARC Challenge	36.2	26.4	41.5
WikiQA	38.4	25.4	**71.2**
AHE	64.6	33.9	68.5

Domain transfer of BiMPM (Wang et al., 2017)

Conclusion

Thank you! Questions?