Introduction to Question Answering

SCADS.AI SUMMER SCHOOL 2022

TECHNISCHE UNIVERSITÄT DARMSTADT

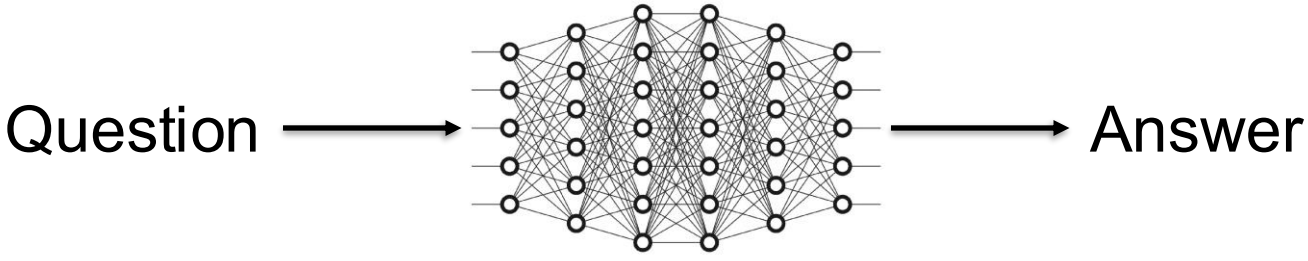
ScaDS.AI DRESDEN LEIPZIG

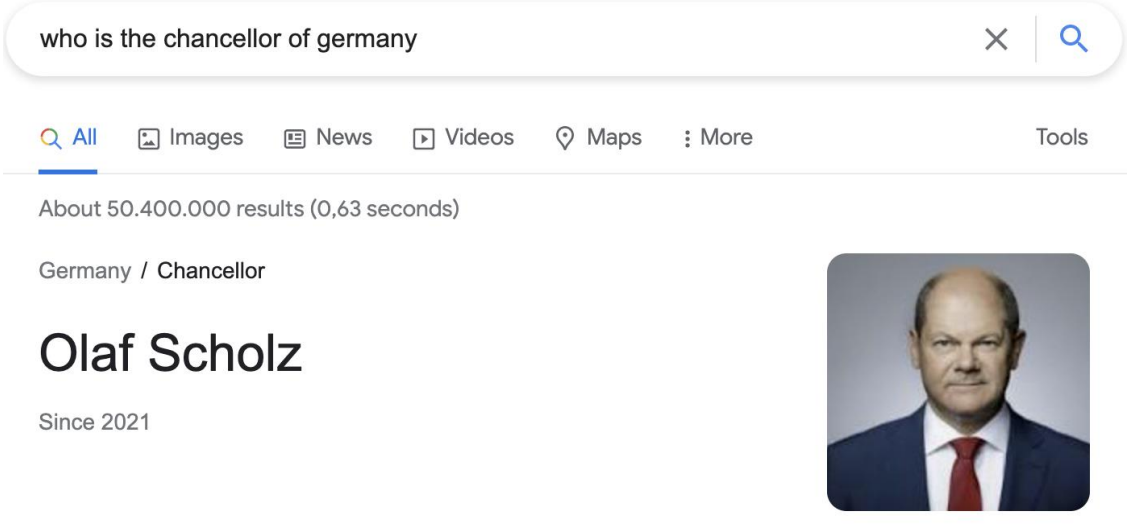
Prof. Dr. Iryna Gurevych Haritz Puerto

Table of Contents

1. INTRODUCTION TO QA

What is QA?


Question
Answer


who is the chancellor of germany
All
Images
News
Videos
Maps
More
Tools
About 50.400.000 results (0,63 seconds)
Germany / Chancellor
Olaf Scholz
Since 2021

What is QA?

Google Assistant

Why is QA important?

- Ideal testbed to evaluate the natural language understanding of AI systems
- Makes the knowledge of the world accessible
- Many other NLP tasks can be modeled as QA

Fact Verification

Claim: The Rodney King riots took place in the most populous county in the USA.

[wiki/Los_Angeles_Riots]
The 1992 Los Angeles riots, also known as the Rodney King riots were a series of riots, lootings, arsons, and civil disturbances that occurred in Los Angeles County, California in April and May 1992.

[wiki/Los_Angeles_County]
Los Angeles County, officially the County of Los Angeles, is the most populous county in the USA.

Verdict: Supported

→ **Q:** Is it true that the Rodney King riots took place in the most populous county in the USA?

Sentiment Analysis

The room was very comfortable.

↓

Q: Is the sentiment positive?
Q: What is the sentiment?

FEVER: a Large-scale Dataset for Fact Extraction and VERification (Thorne et al., NAACL 2018)

QA Types

- Extractive QA (a.k.a. Machine Reading Question Answering)
- Multiple-Choice QA
- Open Domain QA (a.k.a. Open Retrieval)
- Visual QA
- And many others

Extractive QA

The Rhine (Romansh: Rein, German: Rhein, French: le Rhin, Dutch: Rijn) is a European river that begins in the Swiss canton of Graubünden in the southeastern Swiss Alps, forms part of the Swiss-Austrian, Swiss-Liechtenstein border, Swiss-German and then the Franco-German border, then flows through the Rhineland and eventually empties into the North Sea in the Netherlands. The biggest city on the river Rhine is Cologne, Germany with a population of more than 1,050,000 people. It is the second-longest river in Central and Western Europe (after the Danube), at about 1,230 km (760 mi),[note 2][note 1] with an average discharge of about 2,900 m3/s (100,000 cu ft/s).

What is the largest city the Rhine runs through?

Ground Truth Answers: Cologne, Germany | Cologne, Germany | Cologne

Prediction: Cologne, Germany

(Question , Passage) → A

The answer is a contiguous span of the text

* The **passage** is sometimes called *context*

SQuAD: 100,000+ Questions for Machine Comprehension of Text (Rajpurkar et al., EMNLP 2016)

Multiple-Choice QA

Alex spilled the food she just prepared all over the floor and it made a huge mess.

Q What will Alex want to do next?

A
(a) taste the food
(b) mop up ✓
(c) run around in the mess

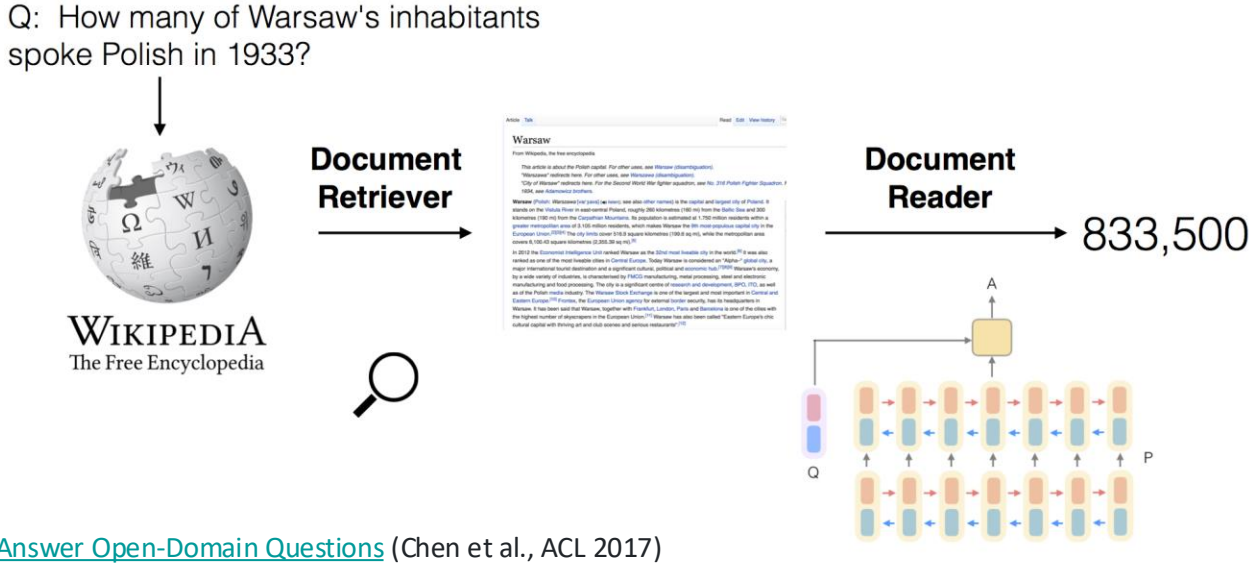
(Question, Passage, Opt_1, …, Opt_k) → A

Social IQa: Commonsense Reasoning about Social Interactions (Sap et al., EMNLP 2019)

Open Domain Question Answering

Question → Answer

Open-domain QA
SQuAD, TREC, WebQuestions, WikiMovies



Reading Wikipedia to Answer Open-Domain Questions (Chen et al., ACL 2017)

Visual QA

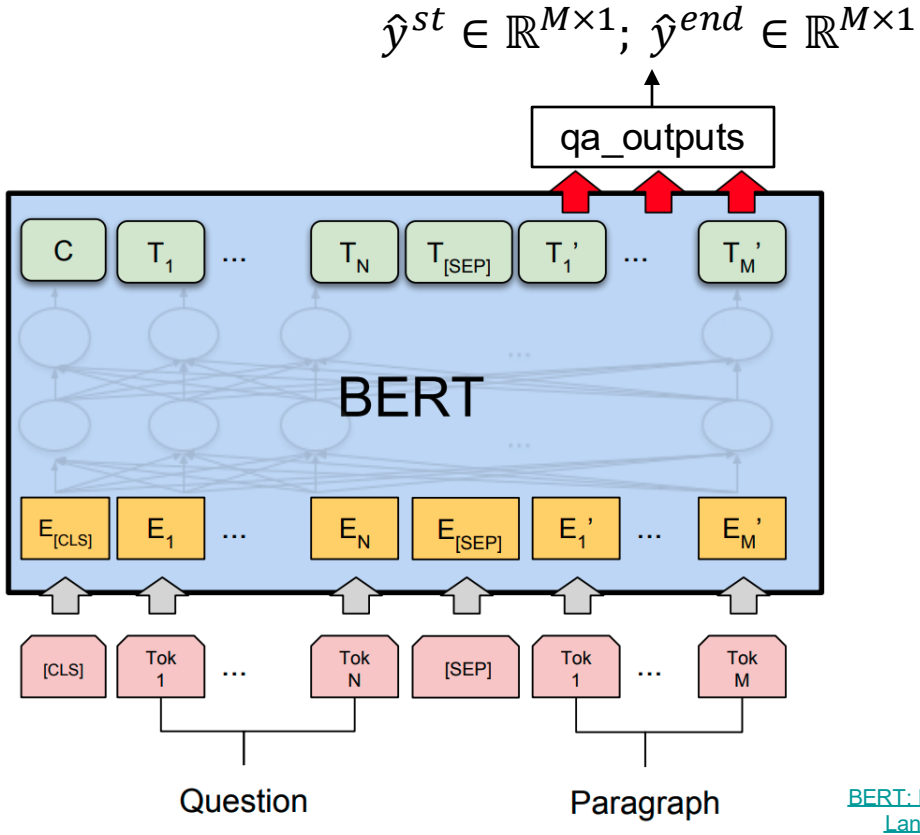
What color are her eyes?
What is the mustache made of?

How many slices of pizza are there?
Is this a vegetarian pizza?

(Q, Img) → A

2. QA MODELS

BERT for QA



BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (Devlin et al., NAACL 2019)

How to Evaluate the Performance?

TECHNISCHE UNIVERSITÄT DARMSTADT

Exact Match (EM)

- Clean the prediction and label
  - Lowercase
  - Remove punctuation
  - Remove articles
  - Fix white spaces
- Prediction == label

F1

- Harmonic mean of the token overlap between the prediction and the label
- F1(“Hello”, “Hello World”) = 0.666
- Token = white-space tokens
- Also, cleans the prediction and label

Is QA solved yet?

HOW GOOD ARE THE MODELS?

QA is not solved yet!

TECHNISCHE UNIVERSITÄT DARMSTADT

Training in one dataset does not generalize to others

Fine-tuned on \ Evaluated on	SQuAD	TriviaQA	NQ	QuAC	NewsQA
SQuAD	**75.6**	46.7	48.7	20.2	41.1
TriviaQA	49.8	**58.7**	42.1	20.4	10.5
NQ	53.5	46.3	**73.5**	21.6	24.7
QuAC	39.4	33.1	33.8	**33.3**	13.8
NewsQA	52.1	38.4	41.7	20.4	**60.1**

Table 3: F1 scores of each fine-tuned model evaluated on each test set

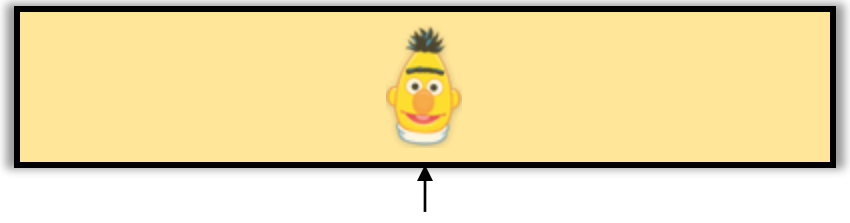
BERT Base uncased

What do Models Learn from Question Answering Datasets? (Sen & Saffari, EMNLP 2020)

QA Generalization

Multi-Dataset Models

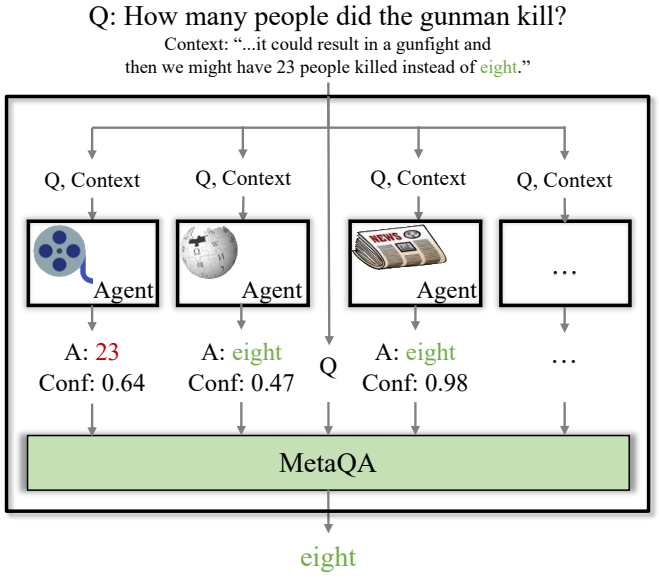
- Train a model on many datasets



UNIFIEDQA: Crossing Format Boundaries with a Single QA System (Khashabi et al., Findings 2020)

Multi-Agent Models

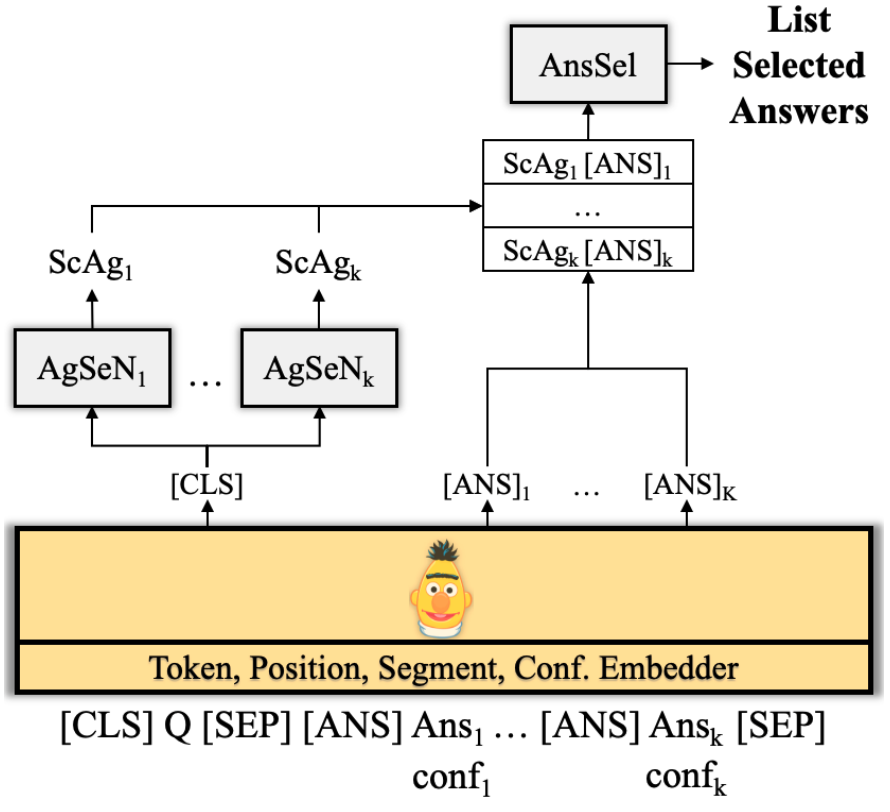
- Combine many models



MetaQA: Combining Expert Agents for Multi-Skill Question Answering (Puerto et al., Arxiv 2021)

MetaQA

- Multi-Task Objective
  - Agent Selection
  - Answer Selection
- Agent Collaboration



Leave-One-Out Ablation

Dataset	TriviaQA	CSQA	HellaSWAG	DuoRC	DROP
MetaQA	80.65	78.40	72.14	51.01	74.96
UnifiedQA	72.34	58.43	36.01	34.65	42.45
OOD MetaQA	**77.26**	**58.75**	**51.94**	**50.64**	22.02
OOD UnifiedQA	69.33	**50.57**	29.35	32.84	**22.30**

…

SQuAD F1 metric

- OOD MetaQA outperforms OOD UnifiedQA by 8.45 in F1
- OOD MetaQA outperforms in-domain UnifiedQA in 4 datasets
- MetaQA outperforms UnifiedQA by 8.89 in F1

3. EXPLAINABILITY IN QA

What is Explainability?

TECHNISCHE UNIVERSITÄT DARMSTADT

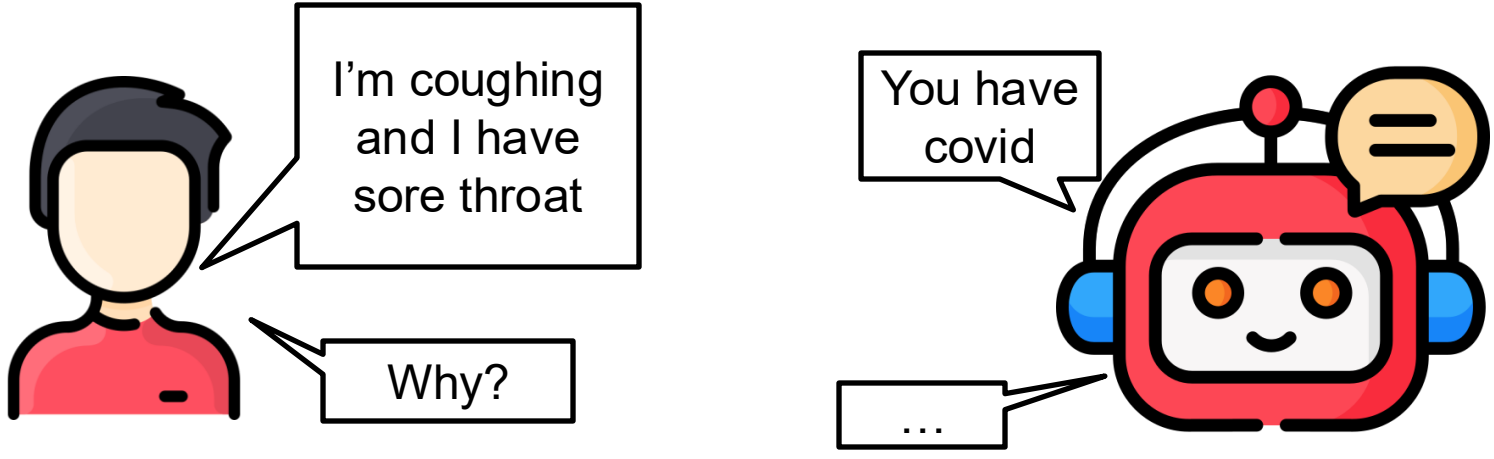
Explainability is stating “how/why” the model gives a prediction

- Fundamental questions in XQA (explainable QA)
  1. Why did the QA system **choose this answer**?
  2. Why did not the QA system answer **something else**?
  3. **When** did the QA system **succeed**?
  4. **When** did the QA system **fail**?
  5. **When** does the QA system give **enough confidence** in the answer that you can trust?
  6. **How** can the QA system **correct** an error?

Shekarpour, S., & Alshargi, F. (2019). A Road-map Towards Explainable Question Answering A Solution for Information Pollution. *arXiv preprint arXiv:1907.02606*.

Why do we need it?

- Allows us to trust the prediction
- Can help us identify wrong predictions
- Sometimes, a prediction alone usually is not useful
  - Eg: Medical QA



Icons from flaticon.com

Explainability in QA

TECHNISCHE UNIVERSITÄT DARMSTADT

Saliency Maps

- Gradient-based Methods
- Attention-based Methods

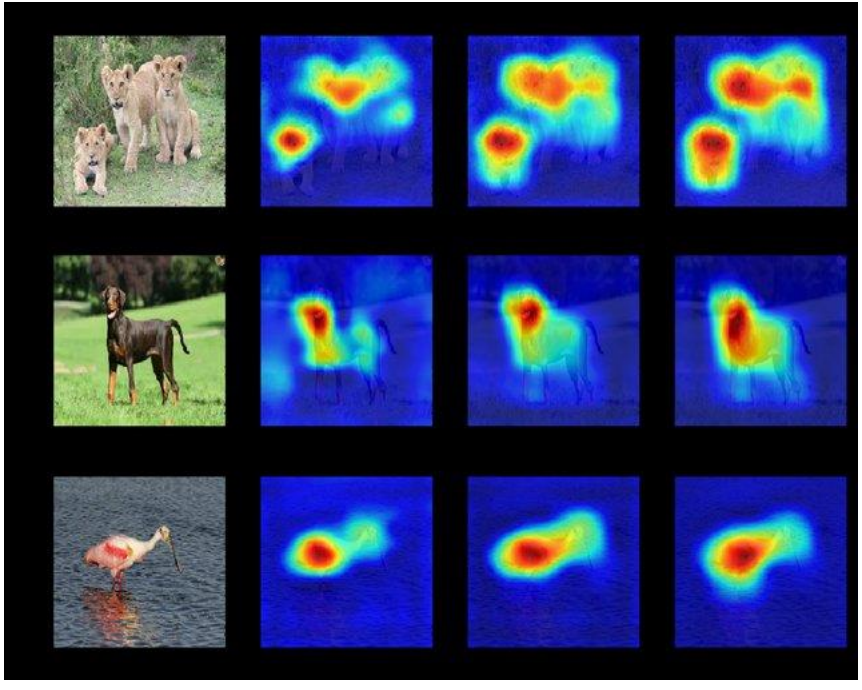
Behavioral Tests

- Minimum Functionality Tests
- Invariance Tests

Saliency Maps

- Inspired by computer vision
- Draw a map that shows the pixels that support the prediction of the class

Applicable to all neural networks



Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Saliency Maps in QA

QUESTION

Who stars in The Matrix ?

Visualizing the top 3 most important words.

PASSAGE

The Matrix is a 1999 science fiction action film written and directed by The Wachowskis , starring Keanu Reeves , Laurence Fishburne , Carrie - Anne Moss , Hugo Weaving , and Joe Pantoliano . It depicts a dystopian future in which reality as perceived by most humans is actually a simulated reality called " the Matrix " : created by sentient machines to subdue the human population , while their bodies ' heat and electrical activity are used as an energy source . Computer programmer " Neo " learns this truth and is drawn into a rebellion against the machines , which involves other people who have been freed from the " dream world . "

Visualizing the top 3 most important words.

https://demo.allennlp.org/reading-comprehension/bidaf-elmo

Saliency Maps, how to compute them?

TECHNISCHE UNIVERSITÄT DARMSTADT

Gradient-based Methods

- Vanilla Gradients [1]
- Integrated Gradients [2]
- SmoothGrad [3]

Attention-based Methods

- Attention Weights
- Scaled Attention [4]

[1] Deep inside convolutional networks: Visualising image classification models and saliency maps (Simonyan et al., arXiv 2013)
[2] Axiomatic attribution for deep networks (Sundararajan et al., PMLR 2017)
[3] Smoothgrad: removing noise by adding noise (Smilkov et al., arXiv 2017)
[4] Is Attention Interpretable? (Serrano & Smith, ACL 2019)

Gradient-based Saliency Maps

TECHNISCHE UNIVERSITÄT DARMSTADT

? What does the gradient tell us?

What weights should be **changed** to **minimize** the **loss**

? What if we use the output **prediction as label** and compute the loss?
Then, what is telling us the gradient?

What weights should be changed to minimize the loss = to maximize the selection of the prediction

Large gradient in a word → changing the word has a big effect on the prediction

Saliency Maps in QA

QUESTION

Who stars in The Matrix ?

Visualizing the top 3 most important words.

PASSAGE

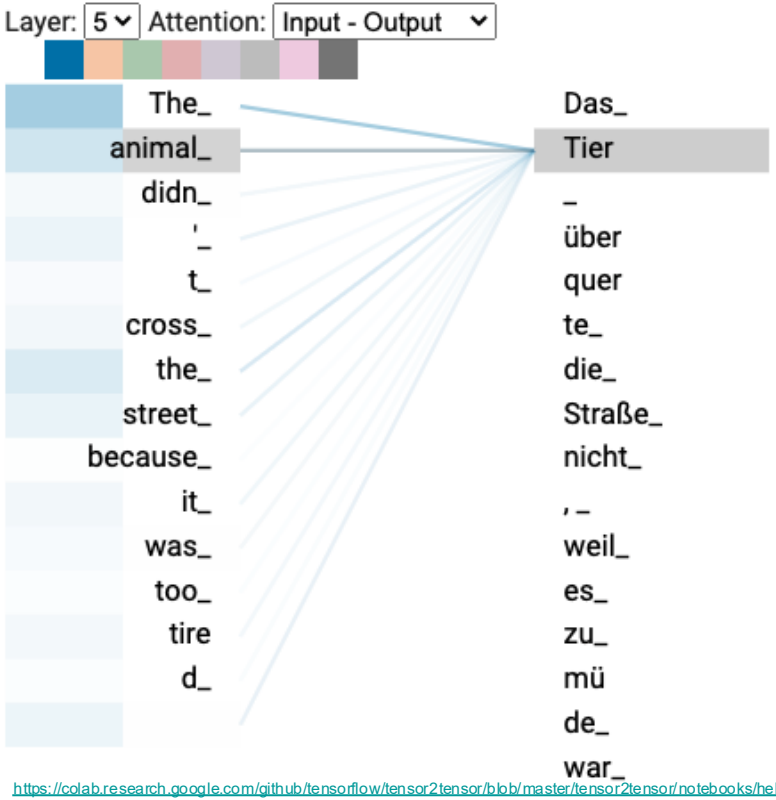
The Matrix is a 1999 science fiction action film written and directed by The Wachowskis , starring Keanu Reeves , Laurence Fishburne , Carrie - Anne Moss , Hugo Weaving , and Joe Pantoliano . It depicts a dystopian future in which reality as perceived by most humans is actually a simulated reality called " the Matrix " : created by sentient machines to subdue the human population , while their bodies ' heat and electrical activity are used as an energy source . Computer programmer " Neo " learns this truth and is drawn into a rebellion against the machines , which involves other people who have been freed from the " dream world . "

Visualizing the top 3 most important words.

https://demo.allennlp.org/reading-comprehension/bidaf-elmo

Attention-based Saliency Maps

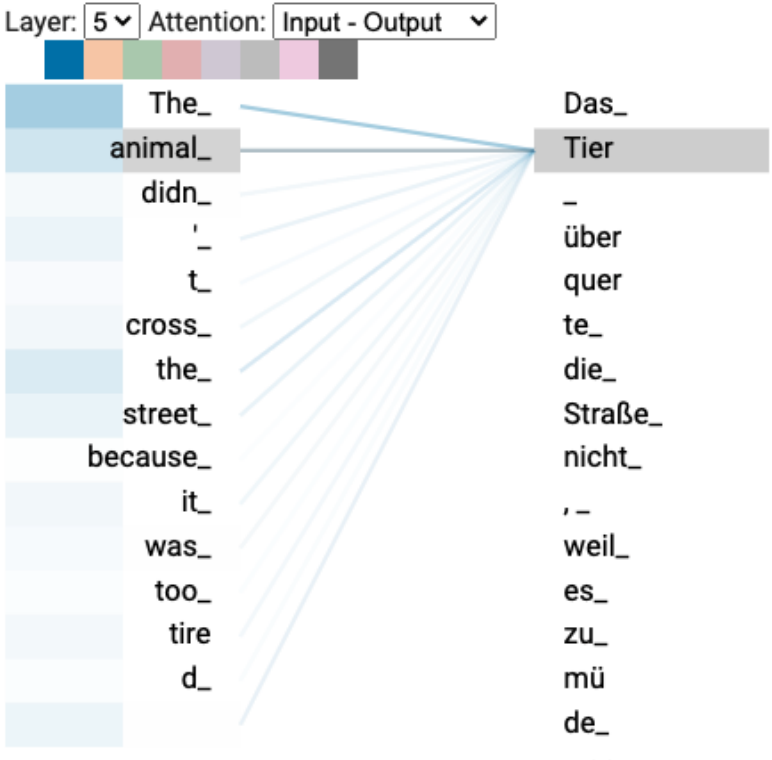
- Attention calculates a distribution over inputs
- It can naturally show the importance of the inputs



https://colab.research.google.com/github/tensorflow/tensor2tensor/blob/master/tensor2tensor/notebooks/hello_t2t.ipynb#scrollTo=OJKU36QAfqOC

Attention-based Saliency Maps

- In QA, use the [CLS] attention weights [1]
- However, attention is highly inconsistent and may not necessarily correspond to importance [1,2]
- Scaled Attention:
  - Attention Scores x Gradient



https://colab.research.google.com/github/tensorflow/tensor2tensor/blob/master/tensor2tensor/notebooks/hello_t2t.ipynb#scrollTo=OJKU36QAfqOC

[1] Is Attention Interpretable? (Serrano & Smith, ACL 2019)
[2] Attention is not Explanation (Jain & Wallace, NAACL 2019)

Behavioral Testing

Validates the input-output behavior w/o knowing the model internals

List of questions and answers that evaluates the behavior of a model

Multiple types of tests

- Minimum Functionality Tests
- Invariance

Animal vs Vehicle

Test's model's ability to understand different animals and vehicles.

Min Func Test test on **Taxonomy**

Question: What vehicle does Victoria have?

Context: Victoria has a snake and a SUV.

Answer: SUV ✓

Prediction: snake ✗

Icons created by Freepik - Flaticon

TECHNISCHE UNIVERSITÄT DARMSTADT

Graph + Text

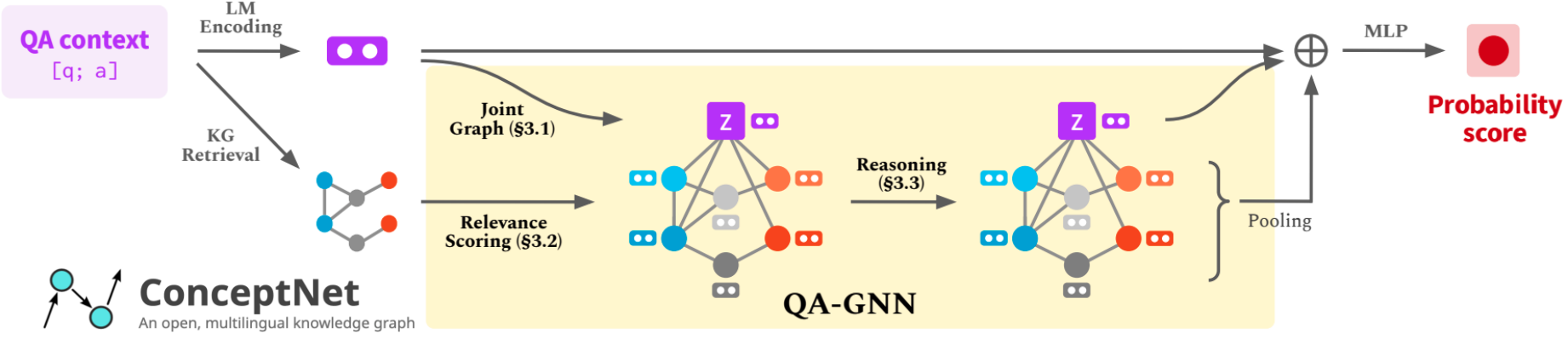
4. NEUROSYMBOLIC QA

UKP

Neurosymbolic QA

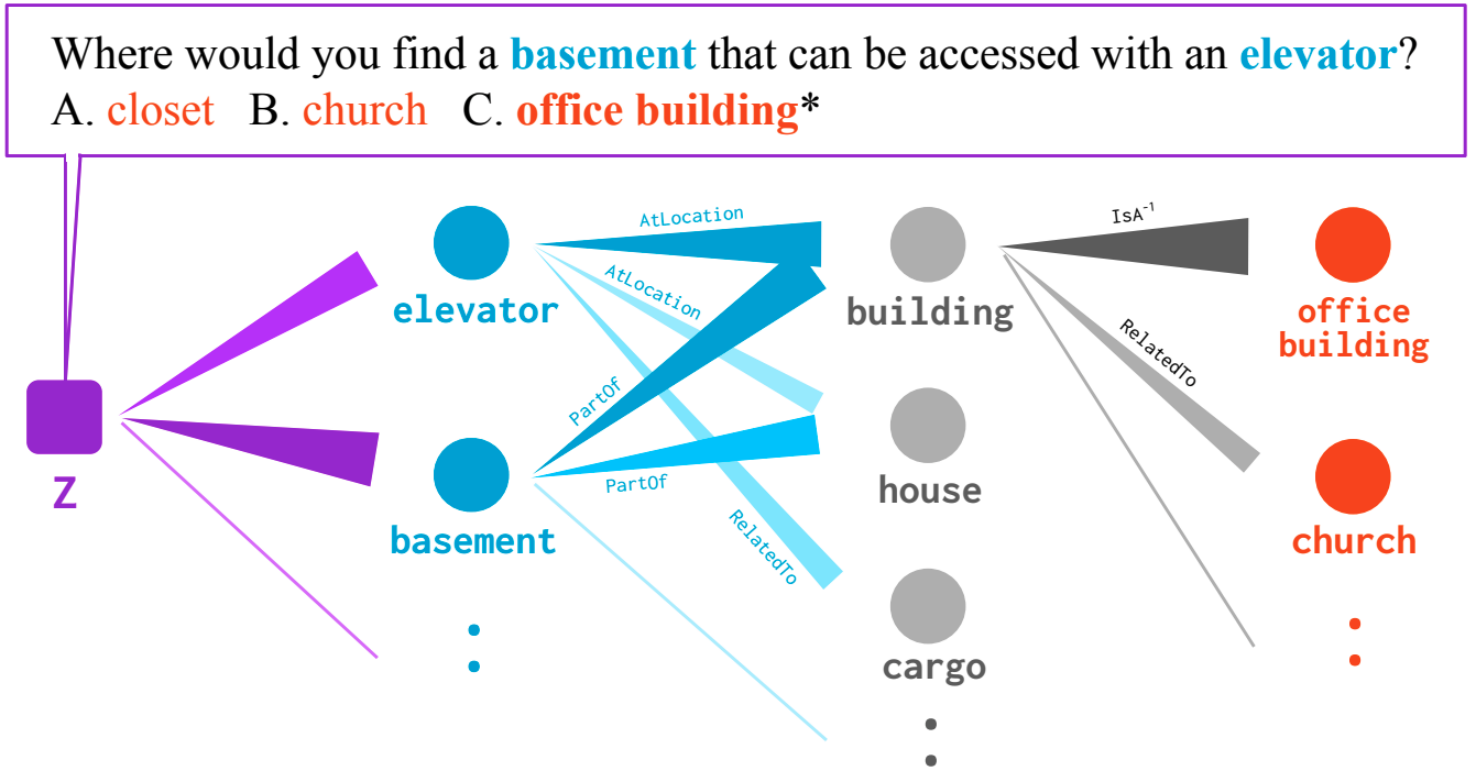
LM:
- Broad coverage of knowledge
- Struggle on structured reasoning

KG:
- Incomplete
- Suited for structured reasoning



QA-GNN: Reasoning with Language Models and Knowledge Graphs for Question Answering (Yasunaga et al., NAACL 2021)

Graph-based Explainability



TECHNISCHE UNIVERSITÄT DARMSTADT

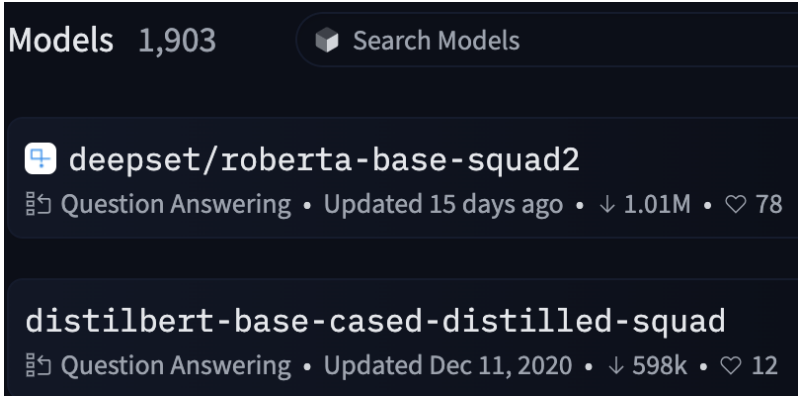
Software for Question Answering Research

5. UKP-SQUARE

UKP

Speed of Research is Overwhelming

- Explosion of QA datasets [1] and models
- Super fast progress
  - Electra outperformed AlBERT on SQuAD 2.0 after 10 days [2]



Icon from flaticon.com

Where to start?

https://huggingface.co/models?pipeline_tag=question-answering&sort=downloads

[1] Rogers, A., Gardner, M., & Augenstein, I. (2021). QA dataset explosion: A taxonomy of nlp resources for question answering and reading comprehension. arXiv preprint arXiv:2107.12708.

[2] https://rajpurkar.github.io/SQuAD-explorer/

QA Frameworks Are Slow for Exploring Research Questions

Learning new APIs

Configurations and setups in your own hardware

Don’t include pretrained models ready to compare

Don’t include explainability and visualization tools

Data icons created by Freepik - Flaticon

UKP-SQuARE: Single Entry Point for QA

TECHNISCHE UNIVERSITÄT DARMSTADT

ACL 2022
22ND – 27TH MAY | 60TH MEETING | DUBLIN

- Common **interface**
- Easy **analysis** of the strengths, weaknesses, and biases
- Common **explainability** methods
- Running on the **browser** (no configurations, no installations)
- Re-use **data sources**
- Easy to **deploy** new **models**
- Dozens of **models available**

https://square.ukp.informatik.tu-darmstadt.de/

Icons created by Freepik - Flaticon

Future Work in UKP-SQuARE

TECHNISCHE UNIVERSITÄT DARMSTADT

Saliency Maps for Explainability

QUESTION

What do robots that resemble humans attempt to do ?

Visualizing the top 3 most important words.

https://demo.allennlp.org/reading-comprehension/bidaf-elmo

Fact Verification

- Multi-modal (Graphs + Text)
- Multi-skill (multiple agents)
- Explainable

Novel Neurosymbolic QA models

ConceptNet BabelNet WIKIDATA

SQuARE as a Multi-Agent QA System

Thank You!

gurevych@ukp.informatik.tu-darmstadt.de

puerto@ukp.informatik.tu-darmstadt.de

ukp-lab.de

square.ukp.informatik.tu-darmstadt.de

TECHNISCHE UNIVERSITÄT DARMSTADT

APPENDIX

UKP

MetaQA Overall Results

Dataset	MetaQA	TWEAC	Exp. Agent	UnifiedQA	MultiQA	Voting
SQuAD	91.98±0.11†	89.09±0.36	92.92	90.81	**93.14±0.18**	90.73
NewsQA	71.71±0.21†	66.86±0.75	**73.68**	65.57	73.59±0.60	66.60
HotpotQA	79.27±0.15†	74.96±0.59	80.60	77.92	**81.68±0.22**	71.71
SearchQA	**81.98±0.25**†‡	80.41±0.22	81.04	81.61	80.45±1.82	68.87
TriviaQA-web	**80.63±0.26**†‡	76.55±0.15	79.34	72.34	77.76±4.15	75.73
NQ	81.20±0.18†	78.06±0.37	81.97	75.58	**82.57±0.30**	72.25
DuoRC	**51.24±0.20**†‡	44.28±0.23	43.77	34.65	46.99±0.15	50.94
QAMR	83.78±0.14†	78.77±0.48	84.00	82.70	**84.62±0.14**	73.07
BoolQ	73.14±0.23†	72.20±0.03	72.17	**81.34**	n.a.	73.88
CSQA	**78.66±0.19**†	77.18±0.18	78.56	58.43	n.a.	68.41
HellaSWAG	73.19±1.01	77.12±0.30	**77.14**	36.01	n.a.	69.33
RACE	84.71±0.05†	83.02±0.27	**84.78**	69.65	n.a.	67.30
SIQA	74.17±0.64	75.39±0.05	**75.44**	61.62	n.a.	70.01
DROP	73.04±1.98†	69.12±0.36	**74.61**	42.45	n.a.	26.18
NarrativeQA	**67.19±0.00**	**67.19±0.00**	**67.19**	57.82	n.a.	**67.19**
HybridQA	**50.94±0.00**	**50.94±0.00**	**50.94**	n.a	n.a	**50.94**

MetaQA Qualitative Analysis

Dataset	Question	In-domain Agent	OOD Agent
DuoRC	Who does Rocky Balboa work for as an enforcer?	Adrian	Tony Gazzo (NewsQA Agent)
TriviaQA-web	Who played the character Mr Chips in the 2002 TV adaptation of Goodbye Mr Chips?	Timothy Carroll	MartinClunes (DuoRC Agent)
SearchQA	This short story, written around 1820, contains the line "If I can but reach that bridge... I am safe"	Legend	Legend of Sleepy Hollow (TriviaQA Agent)