

Question Answering Systems

Melih Barsbey & Merve Ünlü

CMPE 561
Bogazici University

May 10,2017

Introduction

What is Question Answering?

History

Example Use Cases

Approaches to Question Answering

Knowledge-based Question Answering

IR-based Question Answering

Machine Comprehension Oriented Question Answering

Datasets and Success Rates, and Frequently Used Databases

Datasets Used by Different Approaches

In Knowledge-based Question Answering Systems

In IR-based Question Answering Systems

In Machine Comprehension Oriented Systems

Most Commonly Used Databases

QA in Non-English Languages

Systems and Tools Currently Used and Current Applications

Systems and Tools

Applications

Conclusion

Introduction

What is Question Answering(QA)?

Aim Provide accurate answers to questions posed in natural language.

A QA system generally consists of 3 modules :

- ▶ Question Processing
- ▶ Document Processing
- ▶ Answer Processing

A multidisciplinary task: Natural Language Processing, Information Retrieval, Database Management, Cognitive Science

History

Earliest task in NLP around 1960s, generally domain specific

- ▶ **BASEBALL (1961)** *How many games did the Yankees play in July?*
- ▶ **LUNAR (1973)** *How many samples contain Titanium?*
- ▶ **MASQUE (1995)** *What is the salary of each manager?*

With the rise of web, IR-based and Machine Oriented Systems

- ▶ **Webclopedia (2000)** IR-based QA system accessible via the web
- ▶ **Mulder (2001)** fully-automated question-answering system available on the web
- ▶ **AnswerBus (2002)** open-domain QA system based Web IR

Example Use Cases

- ▶ Given the breadth of the area, use cases can be of great variety
- ▶ A doctor, looking for the most common symptoms of meningitis
- ▶ A movie aficionado looking to find the first movie Robert De Niro and Joe Pesci starred together
- ▶ A traveler arriving in a Istanbul, looking for vegan restaurants
- ▶ A computer scientist looking to automatize comprehension of stories

Knowledge-based QA

- ▶ Structured database
- ▶ Simple and small vocabulary
- ▶ If closed-domain for expert users
- ▶ If open-domain for casual users
- ▶ Transform questions into logical forms

When was Ada Lovelace born?

birth-year (Ada Lovelace, ?x)

What is the largest state

$\text{argmax}(\lambda.x.\text{state}(x), \lambda.x.\text{size}(x))$

What states border Texas?

$\lambda x.\text{state}(x) \text{ and borders}(x, \text{texas})$

Table: Logical Forms

Knowledge-based QA

1. **Rule-Based Methods** hand written rules to extract patterns from the questions and answers. Less training data, but expert knowledge is needed. (Ex: Quarc)
2. **Supervised Methods** Training data consists of pairs: a question and its logical form. (Ex: Geoquery)
3. **Semi-supervised Methods** Start with a small labeled data, use supervision distant approaches to boost the dataset.

Evaluating Knowledge-based QA Systems

- ▶ Accuracy is a frequently used method of evaluation in question answering systems
- ▶ F1 score can be used

$$F1 = \frac{2 \times (p \times r)}{(p + r)} \quad (1)$$

- ▶ Sometimes instead of exact match, partial matches can be considered while calculating precision and recall, thus rewarding partially correct answers as well as full ones

Information Retrieval Based QA

- ▶ Also called text-based question answering
- ▶ No structured database
- ▶ Making use of vast amounts of unstructured (or semi-structured) data
- ▶ Most appropriate for factoid questions

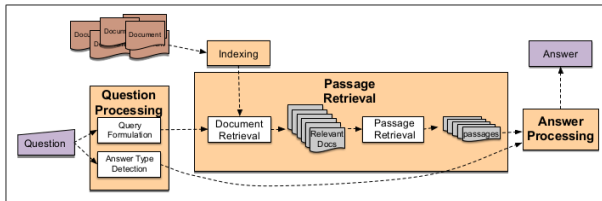


Figure: A generic IR-based QA system [3]

Question Processing in IR-based QA

Two important tasks:

1. **Answer type detection** (a.k.a. question classification)
 - ▶ Deciding what type of answer is wanted
 - ▶ Very important for later stages.
 - ▶ Could be handled with regular expressions or using supervised machine learning
2. **Query formulation**
 - ▶ Deciding on what to feed the information retrieval system
 - ▶ Some parts of the original question removed
 - ▶ Query reformulation or query expansion might be needed

Document/Passage Selection in IR-based QA

- ▶ Finding relevant documents
- ▶ However document is a too coarse unit for ranking results
- ▶ Passages within documents selected
- ▶ Pre-elimination of passages according to the answer type
- ▶ Rank the rest of the passages (most frequently, using supervised machine learning), select the first one

Answer Extraction

- ▶ The system needs to come up with an answer, not just a passage
- ▶ If there's only one named entity of the desired type, it could be returned
- ▶ Else, an n-gram tiling or a pattern-matching algorithm can be used
- ▶ n-gram tiling: Creating an answer using overlapping n-grams

Pattern	Question	Answer
<AP> such as <QP>	What is autism?	“, <u>developmental disorders</u> such as autism”
<QP>, a <AP>	What is a caldera?	“the Long Valley caldera, a <u>volcanic crater</u> 19 miles long”

Figure: A pattern-matching example [3]

Other Systems and Hybrid systems

- ▶ Artificial neural network based models making progress in classical tasks
- ▶ Hybrid system can use various information sources

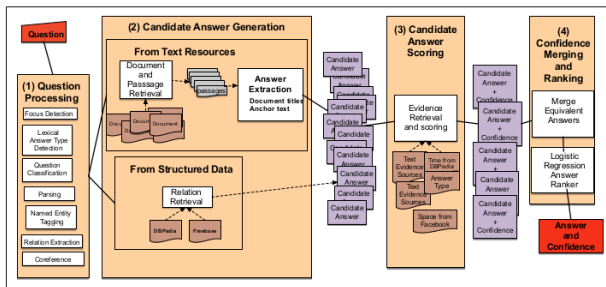


Figure: IBM Watson

Evaluating IR-based QA systems

Two important tasks:

- ▶ One method is using **accuracy** (exact match or F1)
- ▶ Another: **Mean reciprocal rank (MRR)**.
 - ▶ This assumes system returning a list of passages or answers
 - ▶ If the answer is in answer #2, reciprocal rank is $1/2 = 0.5$
 - ▶ This is averaged over all test questions
- ▶ Other methods can factor in confidence in answers, or can produce score if the answer is a list

Machine Comprehension Oriented QA

- ▶ A different way of approaching QA
- ▶ QA not for answering a user-supplied question
- ▶ QA for developing and improving machine comprehension and reasoning.

MCtest Task [6]

James the Turtle was always getting in trouble. Sometimes he'd reach into the freezer and empty out all the food. Other times he'd sled on the deck and get a splinter. His aunt Jane tried as hard as she could to keep him out of trouble, but he was sneaky and got into lots of trouble behind her back.

One day, James thought he would go into town and see what kind of trouble he could get into. He went to the grocery store and pulled all the pudding off the shelves and ate two jars. Then he walked to the fast food restaurant and ordered 15 bags of fries. He didn't pay, and instead headed home.

His aunt was waiting for him in his room. She told James that she loved him, but he would have to start acting like a well-behaved turtle.

After about a month, and after getting into lots of trouble, James finally made up his mind to be a better turtle.

- 1) What is the name of the trouble making turtle?
A) Fries
B) Pudding
C) James
D) Jane
- 2) What did James pull off of the shelves in the grocery store?
A) pudding
B) fries
C) food
D) splinters
- 3) Where did James go after he went to the grocery store?
A) his deck
B) his freezer
C) a fast food restaurant
D) his room
- 4) What did James do after he ordered the fries?
A) went to the grocery store
B) went home without paying
C) ate them
D) made up his mind to be a better turtle

Figure: MCtest Example [6]

Facebook's bAbI Tasks [7]

Task 1: Single Supporting Fact

Mary went to the bathroom.
 John moved to the hallway.
 Mary travelled to the office.
 Where is Mary? A: office

Task 2: Two Supporting Facts

John is in the playground.
 John picked up the football.
 Bob went to the kitchen.
 Where is the football? A: playground

Task 5: Three Argument Relations

Mary gave the cake to Fred.
 Fred gave the cake to Bill.
 Jeff was given the milk by Bill.
 Who gave the cake to Fred? A: Mary
 Who did Fred give the cake to? A: Bill

Task 6: Yes/No Questions

John moved to the playground.
 Daniel went to the bathroom.
 John went back to the hallway.
 Is John in the playground? A: no
 Is Daniel in the bathroom? A: yes

Figure: Example bAbI Tasks [7]

Machine Comprehension Oriented QA

- ▶ Neural network based models dominate
- ▶ Most likely not feedforward ANN's
- ▶ Recurrent neural networks, with longer memory (GRUs, LSTMs)
- ▶ Multimodular systems
- ▶ Attention and memory

Machine Comprehension Oriented QA

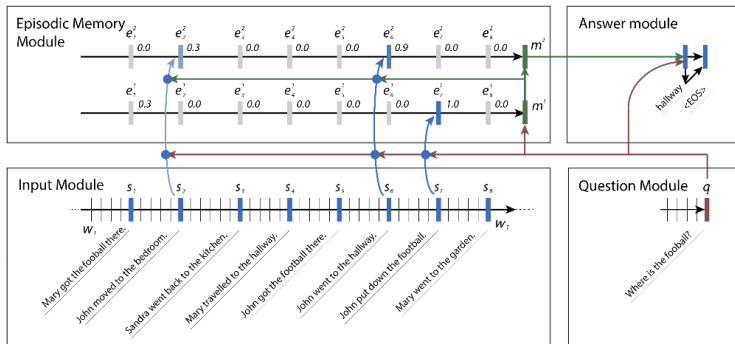


Figure: Dynamic Memory Networks (Initial formulation) [4]

Machine Comprehension Oriented QA

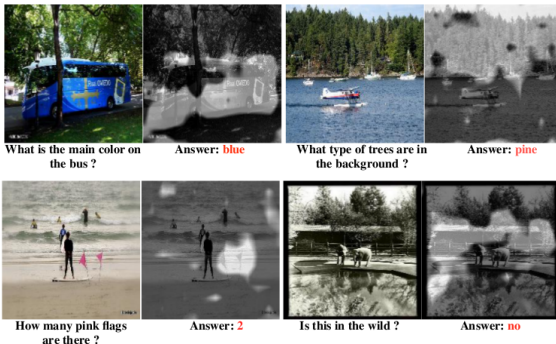


Figure: Visual question answering with improved dynamic memory networks [8]

Evaluating QA systems for machine comprehension

- ▶ Most tasks are hand-crafted and very well defined
- ▶ Strongly supervised
- ▶ Most commonly used metric is exact match accuracy

Common Datasets in Knowledge-based QA Systems

- ▶ **WebQuestions** [1] : 6000 question-answer pairs
 - ▶ Answers provided by Amazon Mechanical Turk workers using Freebase
 - ▶ Current best accuracy (F1) %52,5
- ▶ **Free917** [2]
 - ▶ 917 hand crafted question and answer pairs
 - ▶ Strongly supervised: Allows training for logical parsing.
 - ▶ Best accuracy (F1) as of April 2017: 78.6

Common Datasets in IR-based QA Systems

- ▶ **SQuAD** dataset
- ▶ IR-based methods and machine comprehension
- ▶ More than 100.000 questions

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".

What causes precipitation to fall?

gravity

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?

graupel

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

within a cloud

Figure: SQuAD dataset questions example [5]

Common Datasets in IR-based QA Systems

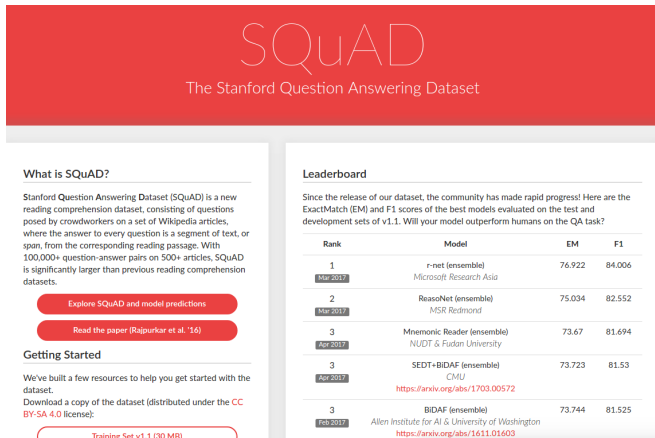


Figure: SQuAD scoreboard [5]

Common Datasets in IR-based QA Systems

- ▶ **WebQA** dataset [9]
- ▶ Derived from Bing queries, annotated by humans
- ▶ Current success rates: $MRR = 0.7069$

Common Datasets in MC-oriented QA Systems

- ▶ **MCtest**
- ▶ Current best performance = 66.2% accuracy

James the Turtle was always getting in trouble. Sometimes he'd reach into the freezer and empty out all the food. Other times he'd sled on the deck and get a splinter. His aunt Jane tried as hard as she could to keep him out of trouble, but he was sneaky and got into lots of trouble behind her back.

One day, James thought he would go into town and see what kind of trouble he could get into. He went to the grocery store and pulled all the pudding off the shelves and ate two jars. Then he walked to the fast food restaurant and ordered 15 bags of fries. He didn't pay, and instead headed home.

His aunt was waiting for him in his room. She told James that she loved him, but he would have to start acting like a well-behaved turtle.

After about a month, and after getting into lots of trouble, James finally made up his mind to be a better turtle.

- 1) What is the name of the trouble making turtle?
A) Fries
B) Pudding
C) James
D) Jane
- 2) What did James pull off of the shelves in the grocery store?
A) pudding
B) fries
C) food
D) splinters
- 3) Where did James go after he went to the grocery store?
A) his deck
B) his freezer
C) a fast food restaurant
D) his room
- 4) What did James do after he ordered the fries?
A) went to the grocery store
B) went home without paying
C) ate them
D) made up his mind to be a better turtle

Figure: MCtest Example [6]

Common Datasets in MC-oriented QA Systems

- ▶ **bAbI** tasks
- ▶ Current best performance = Almost completely solved

<p>Task 1: Single Supporting Fact Mary went to the bathroom. John moved to the hallway. Mary travelled to the office. Where is Mary? A:office</p>	<p>Task 2: Two Supporting Facts John is in the playground. John picked up the football. Bob went to the kitchen. Where is the football? A:playground</p>
<p>Task 5: Three Argument Relations Mary gave the cake to Fred. Fred gave the cake to Bill. Jeff was given the milk by Bill. Who gave the cake to Fred? A: Mary Who did Fred give the cake to? A: Bill</p>	<p>Task 6: Yes/No Questions John moved to the playground. Daniel went to the bathroom. John went back to the hallway. Is John in the playground? A:no Is Daniel in the bathroom? A:yes</p>

Figure: Example bAbI Tasks [7]

Common Datasets in MC-oriented QA Systems

- ▶ **MSCOCO-VQA**
- ▶ Current best performance = 66%



What is the main color on the bus ?



Answer: **blue**



What type of trees are in the background ?



Answer: **pine**



How many pink flags are there ?



Answer: **2**



Is this in the wild ?



Answer: **no**

Most Commonly Used Databases

	DBpedia	Freebase	OpenCyc	Wikidata	YAGO3
Homepage	http://dbpedia.org	http://freebase.com	http://opencyc.org	http://wikidata.org	http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/
Current version	DBpedia 2015-04	continuously updated until Mar 31, 2015*	OpenCyc 4.0	Cont. updated since Oct 2012	YAGO3
Languages	"Main" DBpedia is English (properties etc.), but linked localized versions are available in 125 languages (localized are textual descriptions such as <code>rdfs:label</code> , <code>rdfs:comment</code> , <code>dbpedia-owl:abstract</code> . There are also links to local versions of Wikipedia)	human readable IDs are in English, but every entity and property has an <code>i18n</code> in many languages	English	Almost every language (by community), even dialects	All entity names are from English Wikipedia, some <code>rdfs:label</code> values have different languages
Covered domains	General knowledge	General knowledge, very broad, sometimes deep	Common sense	General knowledge	General knowledge
License (content)	Creative Commons Attribution-ShareAlike 3.0, GNU Free Documentation License	Creative Commons Attribution Only	Creative Commons Attribution 3.0	Creative Commons CC0 1.0 Universal (CC0 1.0) Public Domain Dedication	Creative Commons Attribution 3.0

*Google announced to close Freebase on June 30, 2015. However, currently (July 30, 2015) it is still available.

M. Fahlert et al. / A Comparative Survey of DBpedia, Freebase, OpenCyc, Wikidata, and YAGO

Figure: General Information on Databases

QA in Non-English Languages

Monolingual QA

The questions and answers are in the same language

Cross-lingual QA

The questions and answers are in the different languages

- ▶ Allows users to access information in other languages
- ▶ Recent approaches, translates relevant parts of the question, using machine translation approaches

CLEF-Cross-Language Evaluation Forum an initiative for the evaluation of multilingual IR systems.

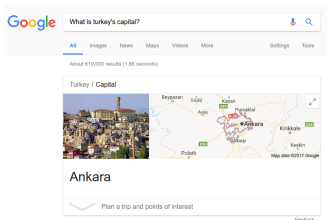
Systems and Tools

- ▶ **YodaQA** an open source Factoid Question Answering system developed in Java, can use structured or unstructured source
- ▶ **Ephyra** open source QA in Java, published under GNU GPL. Oriented toward IR-based systems
- ▶ **Jacana** specially developed for TREC in Java, hosted by Google Code. Use Freebase, developed for knowledge-based and IR based systems.

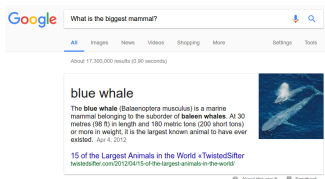
Applications

- ▶ **Personal Assistance** receive an answer or search results accurate to the question asked by user. Voice-controlled like SIRI and Google Now.
- ▶ Mobile Application to access the data like Qme! can handle speech driven questions and returns answers in text format.
- ▶ Using search engines as QA systems

Google Search results



(a) Knowledge-based







(b) IR-based

Figure: Google answers for the questions

Conclusion

- ▶ Ever increasing integration of AI into daily life likely to increase importance of QA even more
- ▶ Likely to stay a vibrant area of research
- ▶ Improvement and increased hybridisation of knowledge-based and IR-based approaches
- ▶ More developments in newer areas such as multimodal, interactive, and machine intelligence oriented question answering

-  Jonathan Berant et al. “Semantic Parsing on Freebase from Question-Answer Pairs.”. In: *EMNLP*. Vol. 2. 5. 2013, p. 6.
-  Qingqing Cai and Alexander Yates. “Large-scale Semantic Parsing via Schema Matching and Lexicon Extension.”. In: *ACL (1)*. 2013, pp. 423–433.
-  Jurafsky D. and James H. Martin. *Speech and Language Processing, 3rd ed.*. Prentice Hall, in press.
-  Ankit Kumar et al. “Ask Me Anything: Dynamic Memory Networks for Natural Language Processing”. In: *CoRR* abs/1506.07285 (2015). URL: <http://arxiv.org/abs/1506.07285>.





Pranav Rajpurkar et al. "SQuAD: 100, 000+ Questions for Machine Comprehension of Text". In: *CoRR* abs/1606.05250 (2016). URL: <http://arxiv.org/abs/1606.05250>.



.MCTest: A Challenge Dataset for the Open-Domain Machine Comprehension of Text. Oct. 2013. URL: <https://www.microsoft.com/en-us/research/publication/mctest-challenge-dataset-open-domain-machine-comprehension-text/>.



Jason Weston et al. "Towards AI-Complete Question Answering: A Set of Prerequisite Toy Tasks". In: *CoRR* abs/1502.05698 (2015). URL: <http://arxiv.org/abs/1502.05698>.

-  Caiming Xiong, Stephen Merity, and Richard Socher. “Dynamic Memory Networks for Visual and Textual Question Answering”. In: *CoRR* abs/1603.01417 (2016). URL: <http://arxiv.org/abs/1603.01417>.
-  Yi Yang, Wen-tau Yih, and Christopher Meek. “WikiQA: A Challenge Dataset for Open-Domain Question Answering.”. In: *EMNLP*. Citeseer. 2015, pp. 2013–2018.