Facts and rules

Facts:

```
(gives daisy milk)
(lives-in daisy pasture)
(has daisy hair)
(eats daisy grass)
...
```

Rules:

```
(Rule 1 (has ?x hair) => (is ?x mammal))
(Rule 2 (is ?x mammal) (has ?x hoofs)
    => (is ?x ungulate))
(Rule 3 (is ?x ungulate) (chews ?x cud) (goes ?x moo)
    => (is ?x cow))
...
```

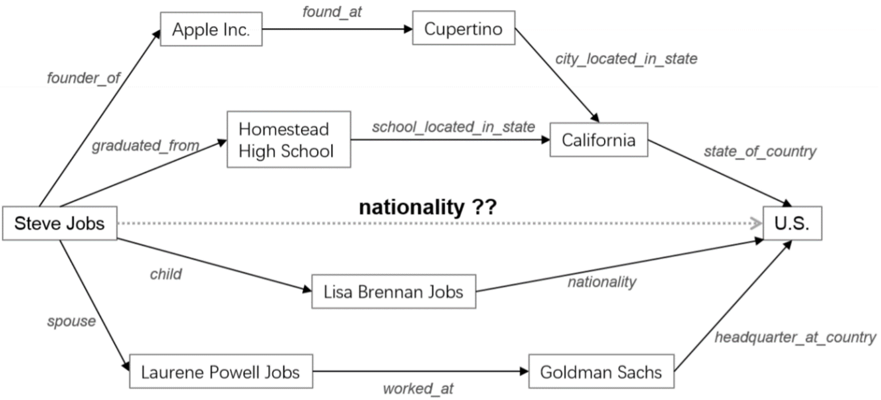
Fernandes et al., A rule-based system proposal to aid in the evaluation..., arXiv:1811.12454

Inference over knowledge graphs

Knowledge graph inference aims to predict relations between entities under supervision of the existing knowledge graph.

Daifeng Li and Andrew Madden, Cascade embedding model for knowledge graph inference and retrieval, Information Processing & Management, 56(6),2019

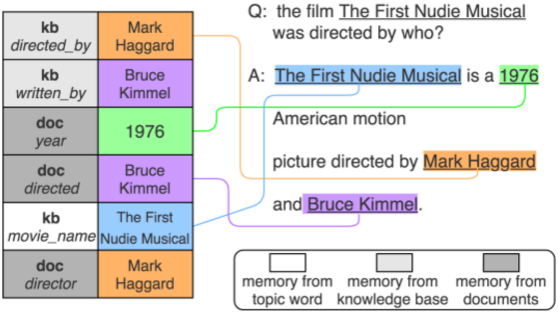
Inference over knowledge graphs



Jiang et al., Attentive Path Combination for Knowledge Graph Completion, ACML 2017

Answer generation

Answer Generation is used when the answer is distributed in databases or knowledgebases, or you want to use more natural answer rather than simple words or phrases.



kb *directed_by*	Mark Haggard
kb *written_by*	Bruce Kimmel
doc *year*	1976
doc *directed*	Bruce Kimmel
kb *movie_name*	The First Nudie Musical
doc *director*	Mark Haggard

Q: the film The First Nudie Musical was directed by who?

A: The First Nudie Musical is a 1976 American motion picture directed by Mark Haggard and Bruce Kimmel.

Yao Fu and Yansong Feng, Natural Answer Generation with Heterogeneous Memory, NAACL-HLT 2018

Content

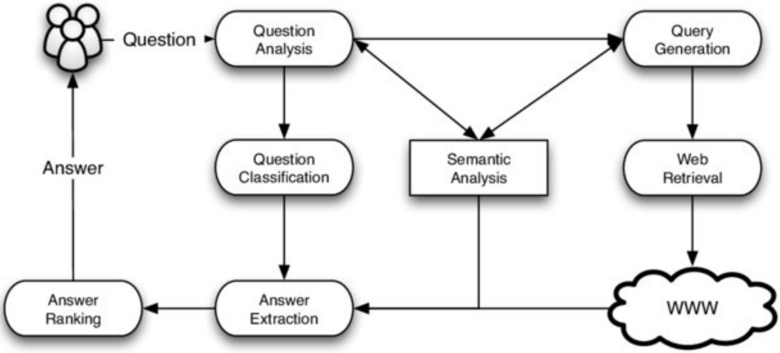
3 Question Answering

- System structure
- Knowledge sources
- Techniques: knowledge representation
- Techniques: question understanding
- Techniques: answer generation
- Open domain QA
- Machine reading comprehension (MRC)

Open domain QA (recap)

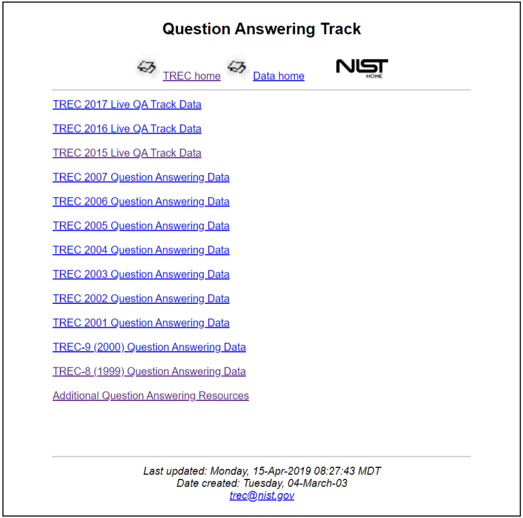
- Open-domain question answering is a category of QA which deals with questions about nearly anything, and can only rely on unstructured data (raw text). On the other hand, these systems usually have much more data available from which to extract the answer.
- The returned answer is in the form of short texts rather than a list of relevant documents (unlike information retrieval systems).
- The system uses a combination of techniques from computational linguistics, information retrieval and knowledge representation for finding answers.

Open domain QA: system structure



Xipeng QiuJiatuo Xu, Interactive Chinese Question Answering System in Medicine Diagnosis, ITME 2008

TREC QA track



Question Answering Track

TREC home | Data home | NIST HOME

- TREC 2017 Live QA Track Data
- TREC 2016 Live QA Track Data
- TREC 2015 Live QA Track Data
- TREC 2007 Question Answering Data
- TREC 2006 Question Answering Data
- TREC 2005 Question Answering Data
- TREC 2004 Question Answering Data
- TREC 2003 Question Answering Data
- TREC 2002 Question Answering Data
- TREC 2001 Question Answering Data
- TREC-9 (2000) Question Answering Data
- TREC-8 (1999) Question Answering Data
- Additional Question Answering Resources

Last updated: Monday, 15-Apr-2019 08:27:43 MDT
Date created: Tuesday, 04-March-03
trec@nist.gov

- Goal:

  Encourage research in information retrieval based on large-scale collections
- Sponsors:
  - NIST
  - ARDA
  - DARPA
- Participants came from research institutes, universities, industries

https://trec.nist.gov/data/qamain.html

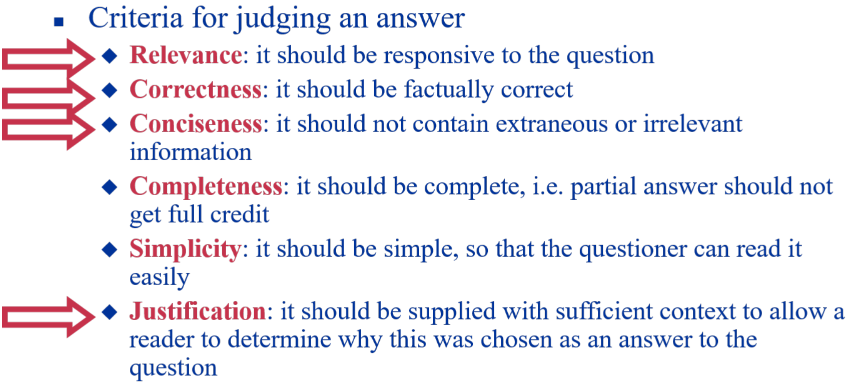
TREC questions

Q-1391: How many feet in a mile?
Q-1057: Where is the volcano Mauna Loa?
Q-1071: When was the first stamp issued?
Q-1079: Who is the Prime Minister of Canada?
Q-1268: Name a food high in zinc.

Q-896: Who was Galileo?
Q-897: What is an atom?

Q-711: What tourist attractions are there in Reims?
Q-712: What do most tourists visit in Reims?
Q-713: What attracts tourists in Reims
Q-714: What are tourist attractions in Reims?

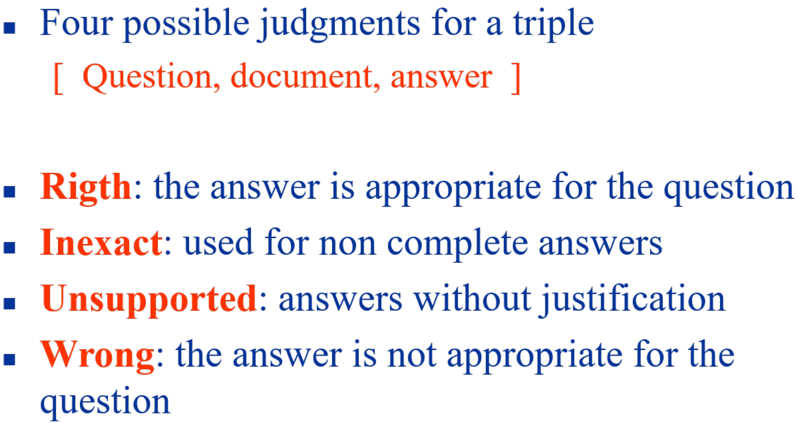
TREC answer assessment



- Criteria for judging an answer
  - **Relevance**: it should be responsive to the question
  - **Correctness**: it should be factually correct
  - **Conciseness**: it should not contain extraneous or irrelevant information
  - **Completeness**: it should be complete, i.e. partial answer should not get full credit
  - **Simplicity**: it should be simple, so that the questioner can read it easily
  - **Justification**: it should be supplied with sufficient context to allow a reader to determine why this was chosen as an answer to the question

Bernardo Magnini, Open Domain Question Answering (slides), RANLP 2005

TREC answer assessment



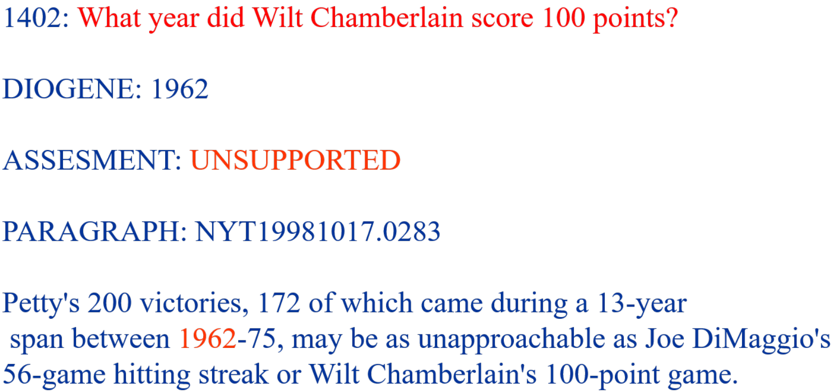
- Four possible judgments for a triple

  [Question, document, answer]

- **Rigth**: the answer is appropriate for the question
- **Inexact**: used for non complete answers
- **Unsupported**: answers without justification
- **Wrong**: the answer is not appropriate for the question

Bernardo Magnini, Open Domain Question Answering (slides), RANLP 2005

TREC answer assessment: examples



1402: What year did Wilt Chamberlain score 100 points?

DIOGENE: 1962

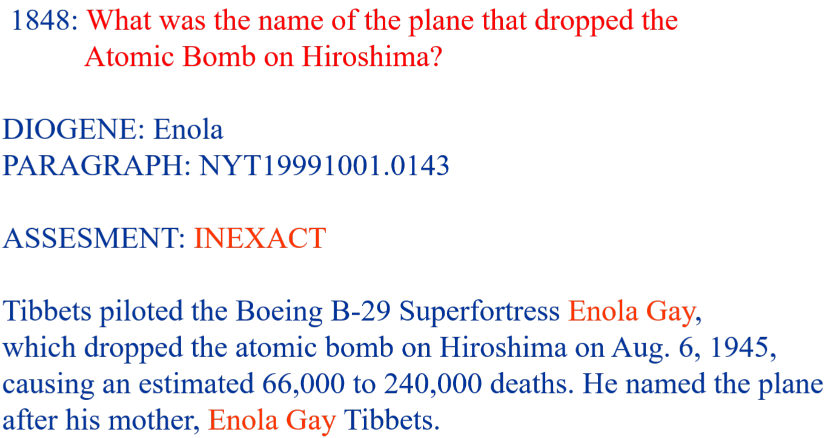
ASSESMENT: UNSUPPORTED

PARAGRAPH: NYT19981017.0283

Petty's 200 victories, 172 of which came during a 13-year span between 1962-75, may be as unapproachable as Joe DiMaggio's 56-game hitting streak or Wilt Chamberlain's 100-point game.

Bernardo Magnini, Open Domain Question Answering (slides), RANLP 2005

TREC answer assessment: examples



1848: What was the name of the plane that dropped the
Atomic Bomb on Hiroshima?

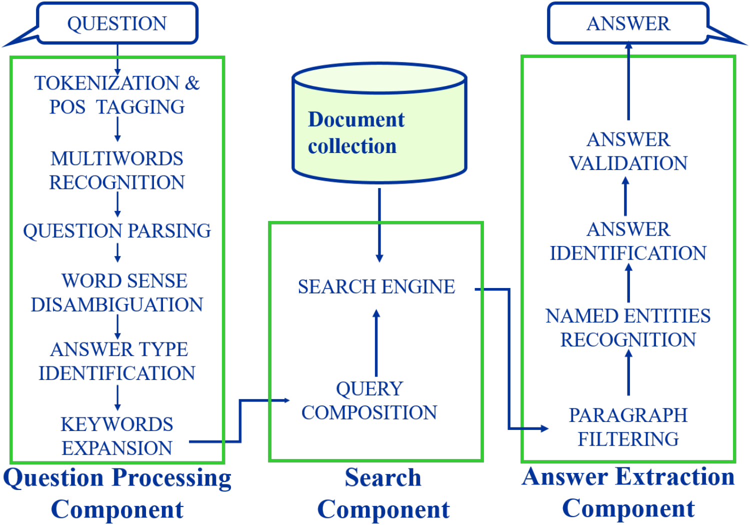
DIOGENE: Enola
PARAGRAPH: NYT19991001.0143

ASSESMENT: INEXACT

Tibbets piloted the Boeing B-29 Superfortress Enola Gay, which dropped the atomic bomb on Hiroshima on Aug. 6, 1945, causing an estimated 66,000 to 240,000 deaths. He named the plane after his mother, Enola Gay Tibbets.

Bernardo Magnini, Open Domain Question Answering (slides), RANLP 2005

TREC QA approaches: Knowledge-Based (1/2)



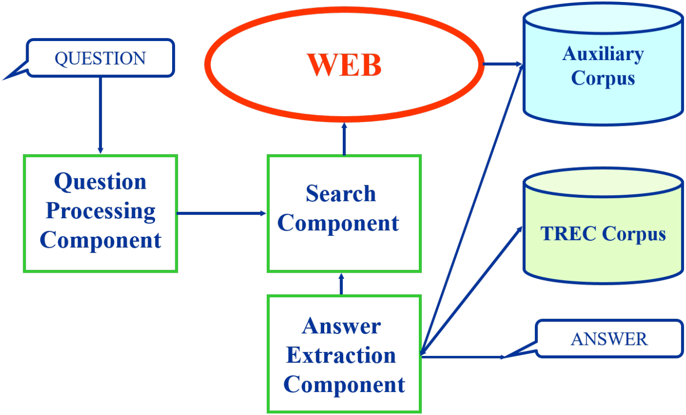
Bernardo Magnini, Open Domain Question Answering (slides), RANLP 2005

TREC QA approaches: Knowledge-Based (2/2)

- **Linguistic-oriented** methodology
  - Determine the answer type from question form
  - Retrieve small portions of documents
  - Find entities matching the answer type category in text snippets
- Majority of systems use a lexicon (usually **WordNet**)
  - To find answer type
  - To verify that a candidate answer is of the correct type
  - To get definitions
- Complex architecture...

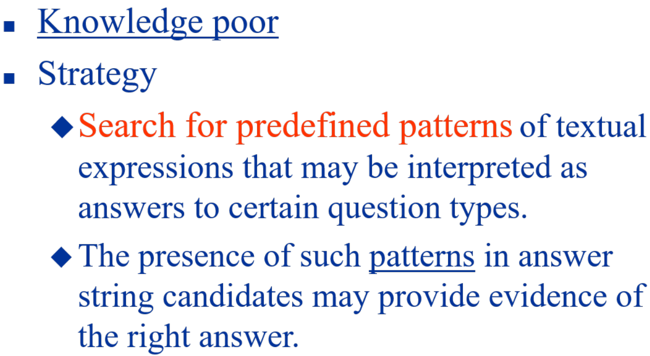
Bernardo Magnini, Open Domain Question Answering (slides), RANLP 2005

TREC QA approaches: Web-Based



Bernardo Magnini, Open Domain Question Answering (slides), RANLP 2005

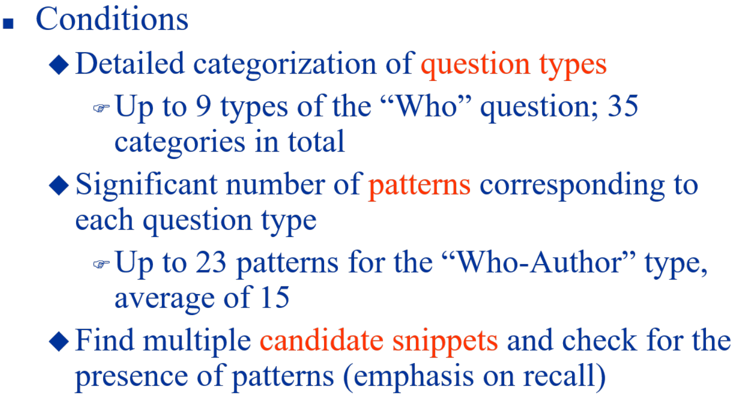
TREC QA approaches: Pattern-Based (1/3)



- Knowledge poor
- Strategy
  - Search for predefined patterns of textual expressions that may be interpreted as answers to certain question types.
  - The presence of such patterns in answer string candidates may provide evidence of the right answer.

Bernardo Magnini, Open Domain Question Answering (slides), RANLP 2005

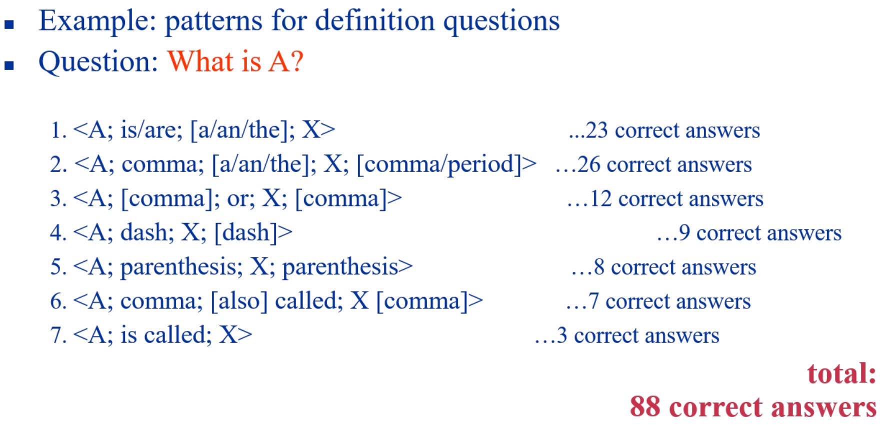
TREC QA approaches: Pattern-Based (2/3)



- Conditions
  - Detailed categorization of question types
    - Up to 9 types of the “Who” question; 35 categories in total
  - Significant number of patterns corresponding to each question type
    - Up to 23 patterns for the “Who-Author” type, average of 15
  - Find multiple candidate snippets and check for the presence of patterns (emphasis on recall)

Bernardo Magnini, Open Domain Question Answering (slides), RANLP 2005

TREC QA approaches: Pattern-Based (3/3)



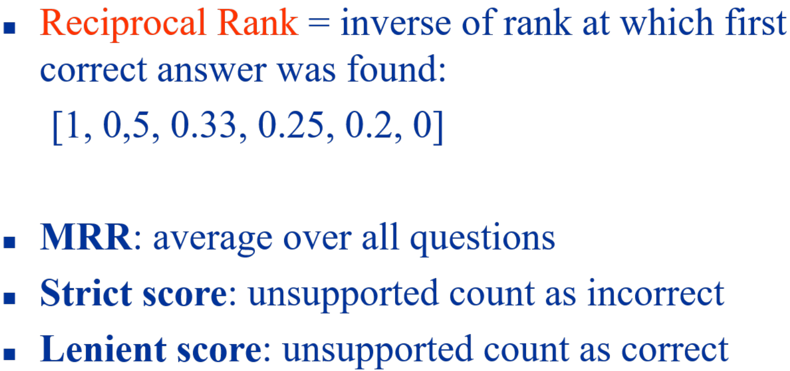
- Example: patterns for definition questions
- Question: What is A?

1. <A; is/are; [a/an/the]; X>	...23 correct answers
2. <A; comma; [a/an/the]; X; [comma/period]>	…26 correct answers
3. <A; [comma]; or; X; [comma]>	…12 correct answers
4. <A; dash; X; [dash]>	…9 correct answers
5. <A; parenthesis; X; parenthesis>	…8 correct answers
6. <A; comma; [also] called; X [comma]>	…7 correct answers
7. <A; is called; X>	…3 correct answers

total:
88 correct answers

Bernardo Magnini, Open Domain Question Answering (slides), RANLP 2005

TREC QA metrics: Mean Reciprocal Rank (MRR)

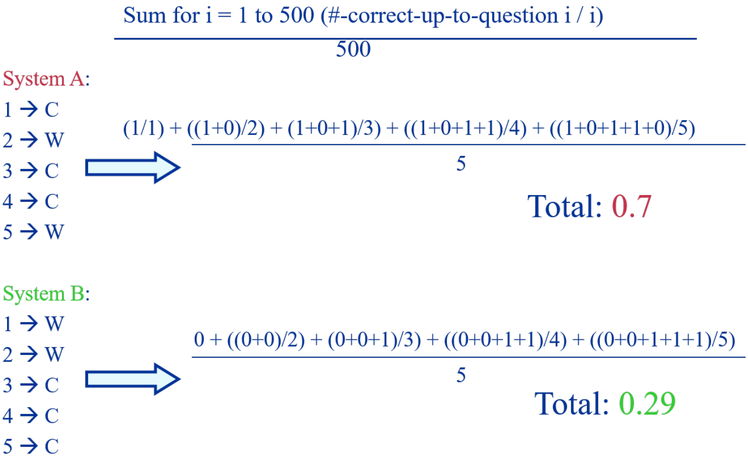


- Reciprocal Rank = inverse of rank at which first correct answer was found:
  [1, 0,5, 0.33, 0.25, 0.2, 0]

- **MRR**: average over all questions
- **Strict score**: unsupported count as incorrect
- **Lenient score**: unsupported count as correct

Bernardo Magnini, Open Domain Question Answering (slides), RANLP 2005

TREC QA metrics: Confidence-Weighted Score (CWS)



$$\frac{\text{Sum for i = 1 to 500 (\#-correct-up-to-question i / i)}}{500}$$

System A:

1 → C
2 → W
3 → C
4 → C
5 → W

$$\frac{(1/1) + ((1+0)/2) + (1+0+1)/3) + ((1+0+1+1)/4) + ((1+0+1+1+0)/5)}{5}$$

Total: 0.7

System B:

1 → W
2 → W
3 → C
4 → C
5 → C

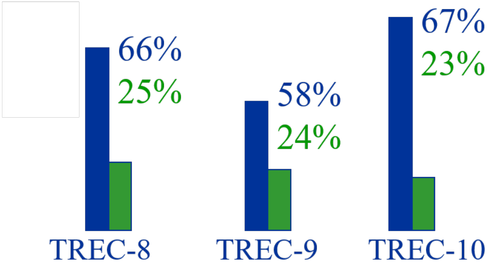
$$\frac{0 + ((0+0)/2) + (0+0+1)/3) + ((0+0+1+1)/4) + ((0+0+1+1+1)/5)}{5}$$

Total: 0.29

Bernardo Magnini, Open Domain Question Answering (slides), RANLP 2005

TREC QA evaluation

- Best result: 67%
- Average over 67 runs: 23%



Bernardo Magnini, Open Domain Question Answering (slides), RANLP 2005

Content

3 Question Answering
- System structure
- Knowledge sources
- Techniques: knowledge representation
- Techniques: question understanding
- Techniques: answer generation
- Open domain QA
- Machine reading comprehension (MRC)

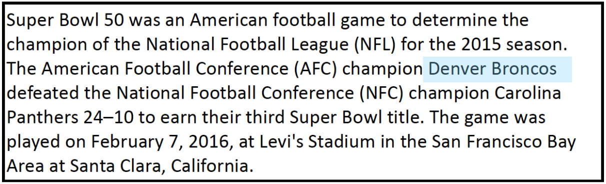
Machine reading comprehension (recap)

- Machine Reading Comprehension (MRC), or Machine Reading (MC), or Machine Comprehension (MC), is the task to read and understand a piece of unstructured text and then answer questions about it.
- MRC is a growing field of research due to its potential in various enterprise applications.
- Although the idea of MRC emerged rather early, only in the past decade, a huge development has been witnessed in this field, including the soar of numbers of corpus (MSMARCO, SQuAD, NewsQA, etc.) and great progress in techniques.

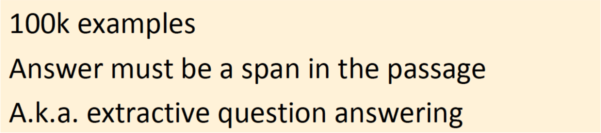
Stanford Question Answering Dataset (SQuAD)

Question: Which team won Super Bowl 50?

Passage



Super Bowl 50 was an American football game to determine the champion of the National Football League (NFL) for the 2015 season. The American Football Conference (AFC) champion Denver Broncos defeated the National Football Conference (NFC) champion Carolina Panthers 24–10 to earn their third Super Bowl title. The game was played on February 7, 2016, at Levi's Stadium in the San Francisco Bay Area at Santa Clara, California.



100k examples

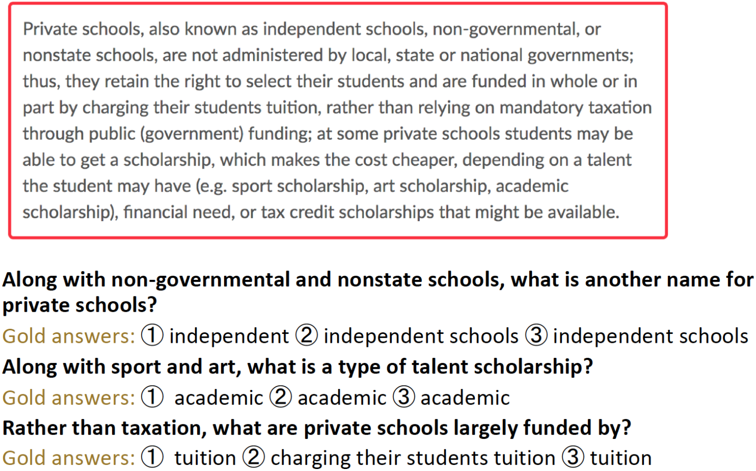
Answer must be a span in the passage

A.k.a. extractive question answering

"SQuAD: 100,000+ questions for machine comprehension of text", Rajpurkar et al., 2016. https://arxiv.org/pdf/1606.05250.pdf

Thomas Lukasiewicz, Advanced Machine Learning: Deep Learning for NLP: Lecture 11: Question Answering, 2019

Stanford Question Answering Dataset (SQuAD)



Private schools, also known as independent schools, non-governmental, or nonstate schools, are not administered by local, state or national governments; thus, they retain the right to select their students and are funded in whole or in part by charging their students tuition, rather than relying on mandatory taxation through public (government) funding; at some private schools students may be able to get a scholarship, which makes the cost cheaper, depending on a talent the student may have (e.g. sport scholarship, art scholarship, academic scholarship), financial need, or tax credit scholarships that might be available.

Along with non-governmental and nonstate schools, what is another name for private schools?

Gold answers: ① independent ② independent schools ③ independent schools

Along with sport and art, what is a type of talent scholarship?

Gold answers: ① academic ② academic ③ academic

Rather than taxation, what are private schools largely funded by?

Gold answers: ① tuition ② charging their students tuition ③ tuition

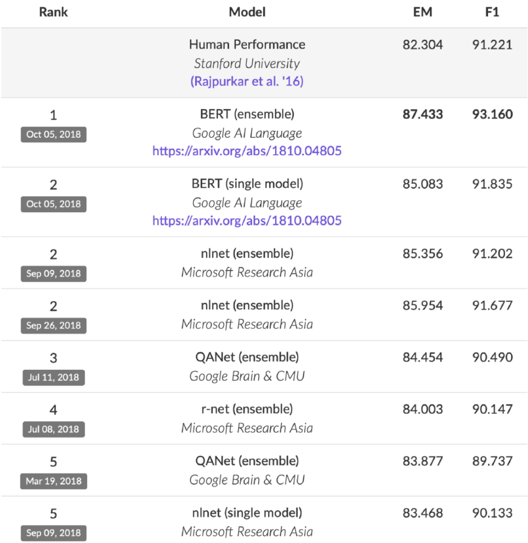
Thomas Lukasiewicz, Advanced Machine Learning: Deep Learning for NLP: Lecture 11: Question Answering, 2019

SQuAD Evaluation, v1.1

- Authors collected 3 gold answers
- Systems are scored on two metrics:
  - Exact match: 1/0 accuracy on whether you match one of the 3 answers
  - F1: Take system and each gold answer as bag of words, evaluate Precision = tp/(tp+fp), Recall = tp/(tp + fn), harmonic mean F1 = 2PR/(P+R) Score is (macro-)average of per-question F1 scores
- F1 measure is seen as more reliable and taken as primary
  - It's less based on choosing exactly the same span that humans chose, which is susceptible to various effects, including line breaks
- Both metrics ignore punctuation and articles (**a**, **an**, **the** only)

Thomas Lukasiewicz, Advanced Machine Learning: Deep Learning for NLP: Lecture 11: Question Answering, 2019

SQuAD v1.1 Leaderboard, 2019-02-07



Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar et al. '16)	82.304	91.221
1 Oct 05, 2018	BERT (ensemble) *Google AI Language* https://arxiv.org/abs/1810.04805	**87.433**	**93.160**
2 Oct 05, 2018	BERT (single model) *Google AI Language* https://arxiv.org/abs/1810.04805	85.083	91.835
2 Sep 09, 2018	nlnet (ensemble) *Microsoft Research Asia*	85.356	91.202
2 Sep 26, 2018	nlnet (ensemble) *Microsoft Research Asia*	85.954	91.677
3 Jul 11, 2018	QANet (ensemble) *Google Brain & CMU*	84.454	90.490
4 Jul 08, 2018	r-net (ensemble) *Microsoft Research Asia*	84.003	90.147
5 Mar 19, 2018	QANet (ensemble) *Google Brain & CMU*	83.877	89.737
5 Sep 09, 2018	nlnet (single model) *Microsoft Research Asia*	83.468	90.133

Thomas Lukasiewicz, Advanced Machine Learning: Deep Learning for NLP: Lecture 11: Question Answering, 2019

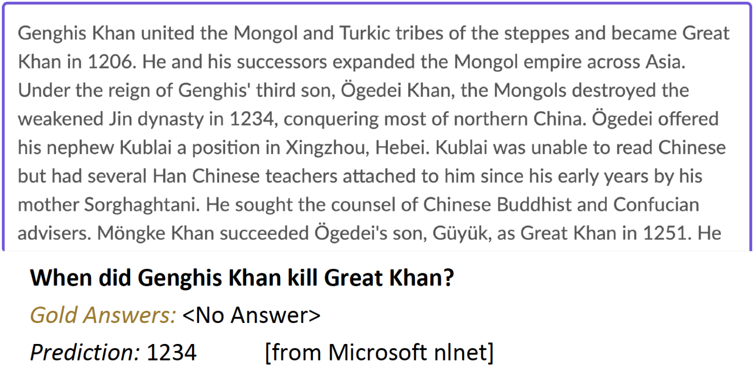
SQuAD 2.0

- A defect of SQuAD 1.0 is that all questions have an answer in the paragraph
- Systems (implicitly) rank candidates and choose the best one
- You don't have to judge whether a span answers the question
- In SQuAD 2.0, 1/3 of the training questions have no answer, and about 1/2 of the dev/test questions have no answer
    - For NoAnswer examples, NoAnswer receives a score of 1, and any other response gets 0, for both exact match and F1
- Simplest system approach to SQuAD 2.0:
    - Have a threshold score for whether a span answers a question
- Or you could have a second component that confirms answering
    - Like Natural Language Inference (NLI) or "Answer validation"

https://rajpurkar.github.io/SQuAD-explorer/

Thomas Lukasiewicz, Advanced Machine Learning: Deep Learning for NLP: Lecture 11: Question Answering, 2019

SQuAD 2.0 Example



Genghis Khan united the Mongol and Turkic tribes of the steppes and became Great Khan in 1206. He and his successors expanded the Mongol empire across Asia. Under the reign of Genghis' third son, Ögedei Khan, the Mongols destroyed the weakened Jin dynasty in 1234, conquering most of northern China. Ögedei offered his nephew Kublai a position in Xingzhou, Hebei. Kublai was unable to read Chinese but had several Han Chinese teachers attached to him since his early years by his mother Sorghaghtani. He sought the counsel of Chinese Buddhist and Confucian advisers. Möngke Khan succeeded Ögedei's son, Güyük, as Great Khan in 1251. He

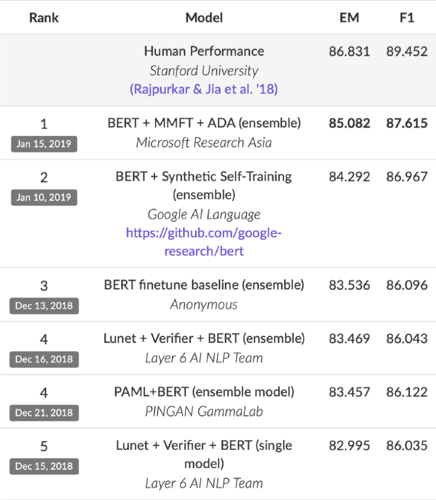
When did Genghis Khan kill Great Khan?

Gold Answers: <No Answer>

Prediction: 1234 [from Microsoft nlnet]

Thomas Lukasiewicz, Advanced Machine Learning: Deep Learning for NLP: Lecture 11: Question Answering, 2019

SQuAD 2.0 leaderboard, 2019-02-07

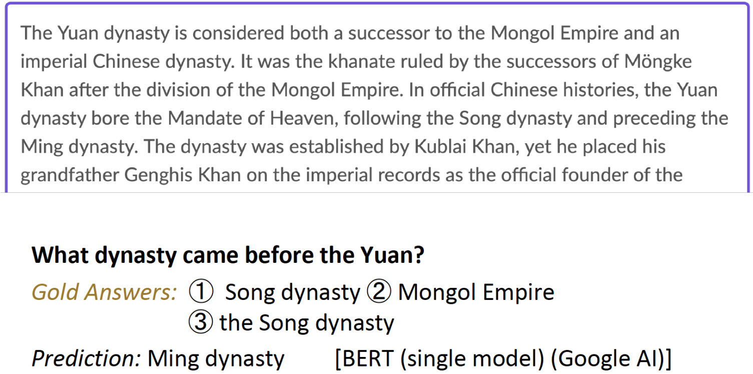


Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Jan 15, 2019	BERT + MMFT + ADA (ensemble) *Microsoft Research Asia*	**85.082**	**87.615**
2 Jan 10, 2019	BERT + Synthetic Self-Training (ensemble) *Google AI Language* https://github.com/google-research/bert	84.292	86.967
3 Dec 13, 2018	BERT finetune baseline (ensemble) *Anonymous*	83.536	86.096
4 Dec 16, 2018	Lunet + Verifier + BERT (ensemble) *Layer 6 AI NLP Team*	83.469	86.043
4 Dec 21, 2018	PAML+BERT (ensemble model) *PINGAN GammaLab*	83.457	86.122
5 Dec 15, 2018	Lunet + Verifier + BERT (single model) *Layer 6 AI NLP Team*	82.995	86.035

Thomas Lukasiewicz, Advanced Machine Learning: Deep Learning for NLP: Lecture 11: Question Answering, 2019

Example

Good systems are great, but still basic NLU errors:



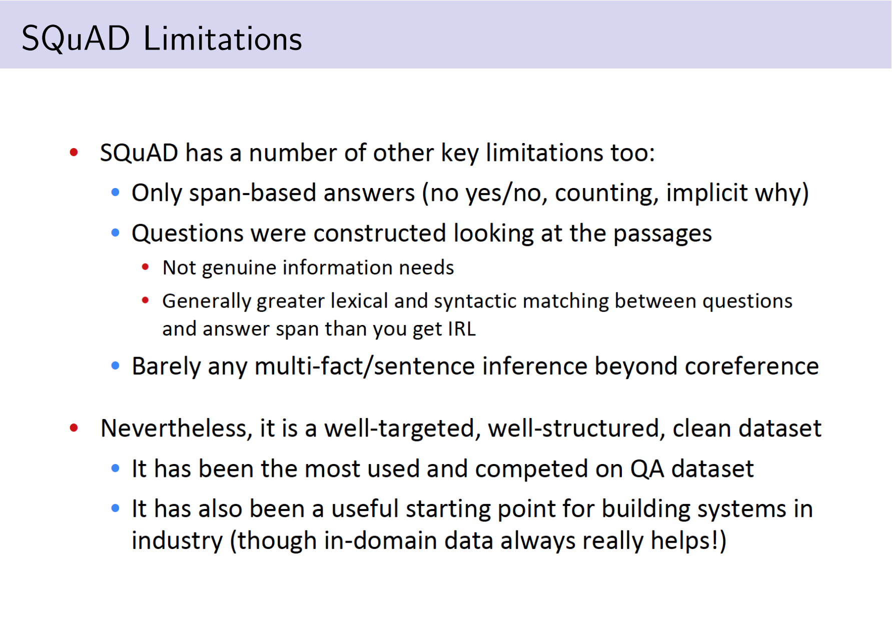
The Yuan dynasty is considered both a successor to the Mongol Empire and an imperial Chinese dynasty. It was the khanate ruled by the successors of Möngke Khan after the division of the Mongol Empire. In official Chinese histories, the Yuan dynasty bore the Mandate of Heaven, following the Song dynasty and preceding the Ming dynasty. The dynasty was established by Kublai Khan, yet he placed his grandfather Genghis Khan on the imperial records as the official founder of the

What dynasty came before the Yuan?

Gold Answers: ① Song dynasty ② Mongol Empire
③ the Song dynasty

Prediction: Ming dynasty [BERT (single model) (Google AI)]

Thomas Lukasiewicz, Advanced Machine Learning: Deep Learning for NLP: Lecture 11: Question Answering, 2019



SQuAD Limitations

- SQuAD has a number of other key limitations too:
  - Only span-based answers (no yes/no, counting, implicit why)
  - Questions were constructed looking at the passages
    - Not genuine information needs
    - Generally greater lexical and syntactic matching between questions and answer span than you get IRL
  - Barely any multi-fact/sentence inference beyond coreference
- Nevertheless, it is a well-targeted, well-structured, clean dataset
  - It has been the most used and competed on QA dataset
  - It has also been a useful starting point for building systems in industry (though in-domain data always really helps!)

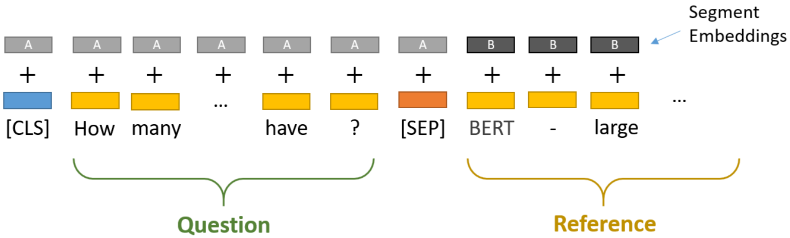
Thomas Lukasiewicz, Advanced Machine Learning: Deep Learning for NLP: Lecture 11: Question Answering, 2019

Fine-tuning BERT for SQuAD (1/6)

- To feed a QA task into BERT, we pack both the question and the reference text into the input.
- The two pieces of text are separated by the special [SEP] token.
- BERT also uses “Segment Embeddings” to differentiate the question from the reference text. These are simply two embeddings (for segments “A” and “B”) that BERT learned, and which it adds to the token embeddings before feeding them into the input layer.

Chris McCormick, Question Answering with a Fine-Tuned BERT (Blog)

Fine-tuning BERT for SQuAD (2/6)



Question: How many parameters does BERT-large have?

Reference Text: BERT-large is really big... it has 24 layers and an embedding size of 1,024, for a total of 340M parameters! Altogether it is 1.34GB, so expect it to take a couple minutes to download to your Colab instance.

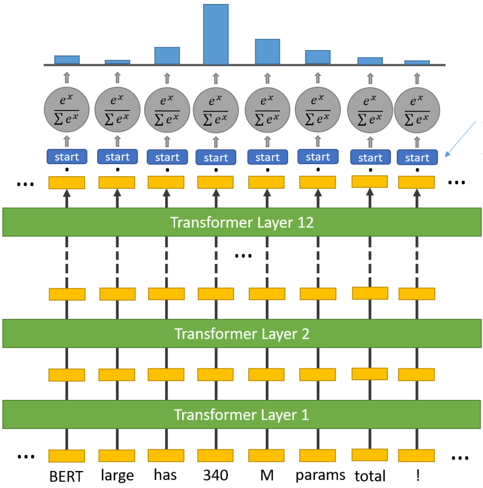
Chris McCormick, Question Answering with a Fine-Tuned BERT (Blog)

Fine-tuning BERT for SQuAD (3/6)

- For every token in the text, we feed its final embedding into the **start token classifier**.
- The **start token classifier** only has a single set of weights (represented by the blue "start" rectangle in the following illustration) which it applies to every word.
- After taking the dot product between the output embeddings and the 'start' weights, we apply the softmax activation to produce a probability distribution over all of the words.
- Whichever word has the highest probability of being the start token is the one that we pick.

Chris McCormick, Question Answering with a Fine-Tuned BERT (Blog)

Fine-tuning BERT for SQuAD (4/6)



This length 768 vector is the ***weights*** *for the start token classifier.*
The ***same weights*** *are applied to* ***every position****.*

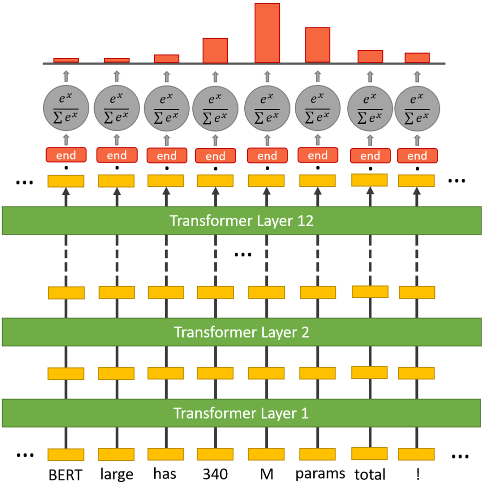
Chris McCormick, Question Answering with a Fine-Tuned BERT (Blog)

Fine-tuning BERT for SQuAD (5/6)

- We repeat this process for the end token – we have a separate weight vector this: the **end token classifier** (represented by the red “start” rectangle in the following illustration).
- The parameters for the **start token classifier** and the **end token classifier** are tuned by the SQuAD training data.
- Note: It’’’s a little naive to pick the highest scores for start and end–what if it predicts an end word that’s before the start word?! The correct implementation is to pick the highest total score for which end >= start.

Chris McCormick, Question Answering with a Fine-Tuned BERT (Blog)

Fine-tuning BERT for SQuAD (6/6)



Chris McCormick, Question Answering with a Fine-Tuned BERT (Blog)

Other QA datasets

- We can see the performance of MRC on SQuAD datasets achieves a level comparable to humans, but it does not mean MRC is a solved problem.
- Many recent research has shown that such MRC systems does not really understand the text, and can be easily attached.
- More QA datasets are constructed by researchers:

bAbI (Facebook)
CoQA
MS MARCO
RACE
CNN / Daily Mail (DeepMind)
HotpotQA
NewsQA
DuReader

Content

1. Introduction to Conversational AI
2. A brief history of QA and dialog systems
3. Question Answering
4. Dialog systems (chatbots)

Content

4 Dialog systems (chatbots)

- Introduction to dialog systems
- Task-oriented dialog system
- Chitchat dialog system

Dialog systems as new human-machine interface

- Unlike QA systems, a dialog system (also called a chatbot) can interact with users with multi-turn conversations.
- Difference between a dialog system and a QA system:
  - A dialog system should understand the conversional context, while a QA system not.
  - A dialog system is expected to handle complex tasks rather than answer a single questions.
  - A dialog system is expected to interact with users friendly rather than simply give answers.

Task-oriented vs. chitchat dialog systems

Two main categories of dialog systems:

- Task-oriented (or goal-oriented) dialog systems
  - Help users to complete certain types of tasks.
  - Tasks (or domain knowledge) should be given in advanced, usually as a set of pre-defined intentions and slots.
  - Dialog sessions: the shorter is the better.
- Chitchat dialog systems (or social chatbots, social bots).
  - Chitchat with users on unrestricted topics.
  - Maximize user engagement by generating enjoyable and more human-like conversations.
  - Emotional conversation and personality is welcome.
  - Dialog sessions: the longer is the better.

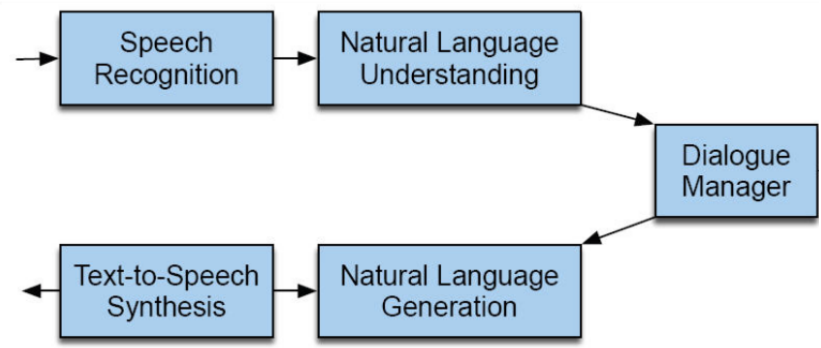
Task-oriented vs. chitchat dialog systems

- Some commercial systems tend to combine the abilities of both categories.
- For example, a voice assistant for mobile phones is able to:
  - set an alarm;
  - make a schedule;
  - tell the whether;
  - play a music;
  - turn on/off wifi;
  - configure your mobile phone;
  - send a message to a person in the contacts;
  - ... etc., and
  - chitchat with the user freely.

Content

4 Dialog systems (chatbots)
- Introduction to dialog systems
- Task-oriented dialog system
- Chitchat dialog system

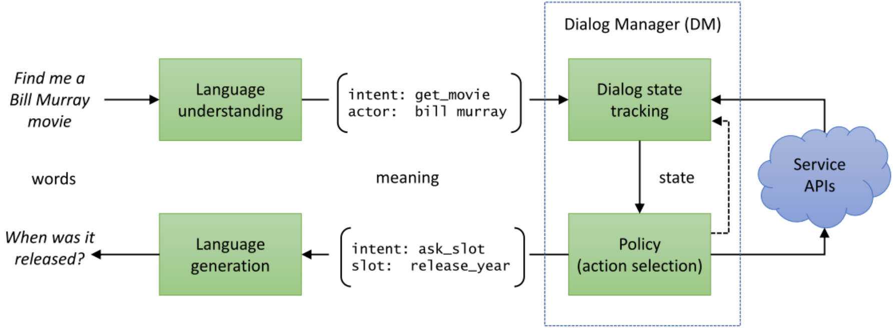
Dialog system structure



System components

- ASR: Automatic Speech Recognition
- NLU: Natural Language Understanding
- DM: Dialog Management
- NLG: Natural Language Generation
- TTS: Text to Speech

Dialog system structure: a closer look



Jianfeng Gao, Michel Galley, Neural Approaches to Conversational AI (slides), ICML 2019

System components - more details

- NLU: Natural Language Understanding
  - **Intent detection**
  - **Slot filling**
- DST: Dialog State Tracking
  - Tracking the **dialog states** according to the users' natural language input and the dialog history
- DP: Dialog Policy
  - Select appropriate **dialog acts** (or **actions**) to response the user input
- NLG: Natural Language Generation
  - Implement the **dialog acts** (or **actions**) with natural language and generate the system **response**