Natural Language Processing
Lecture 10 Conversational AI;
Question Answering; Dialog Systems

Qun Liu, Valentin Malykh
Huawei Noah's Ark Lab

Spring 2022
A course delivered at KFU, Kazan

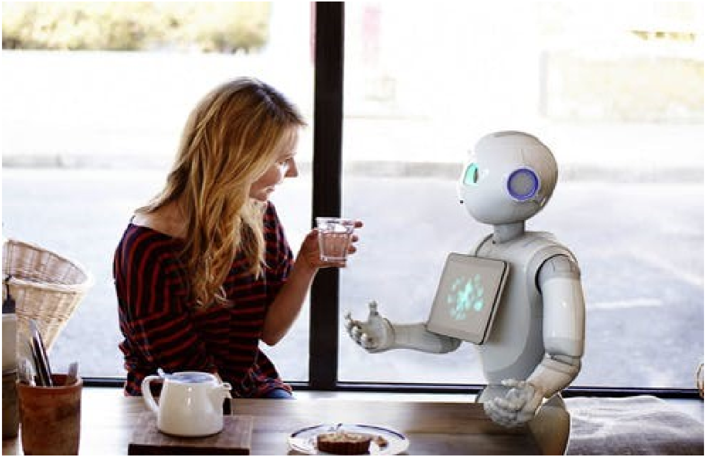
Content

Content

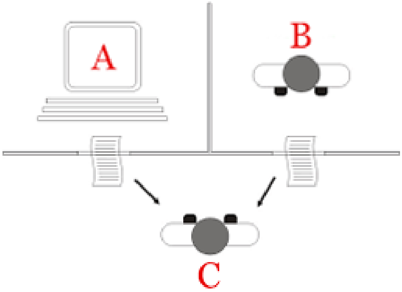
Holy grails of NLP (Recap)

- Accurate machine translation between human languages
- Free conversation between humans and computers

Free human machine conversation (Recap)



Turing test (Recap)



By Juan Alberto Sánchez Margallo, CC BY 2.5, from Wikipedia

Classifications of conversional systems

- Question Answering (QA) Systems
  - Single turn conversation: no dialog context is involved
  - The objective is to answer user's questions
- Dialog Systems
  - Multi-turn conversation: dialog context is involved
  - Diverse objectives: task completion, chitchat, QA
- Multimodal QA/Dialog Systems
  - Additional modality is involved
  - Single turn (Visual QA) or multi-turn (Visual Dialog)
  - Ojbective: conversation around the information given in the additional modalities

Content

Content

2. A brief history of QA and dialog systems
   - Early QA systems
   - Big data era: open domain QA
   - Neural era: machine comprehension and dialog systems

Early QA systems

- BASEBALL(1961) and LUNAR(1971-1973)
- SHRDLU(1968-1970)
- Eliza (1964)
- Expert Systems (1970s-1980s)

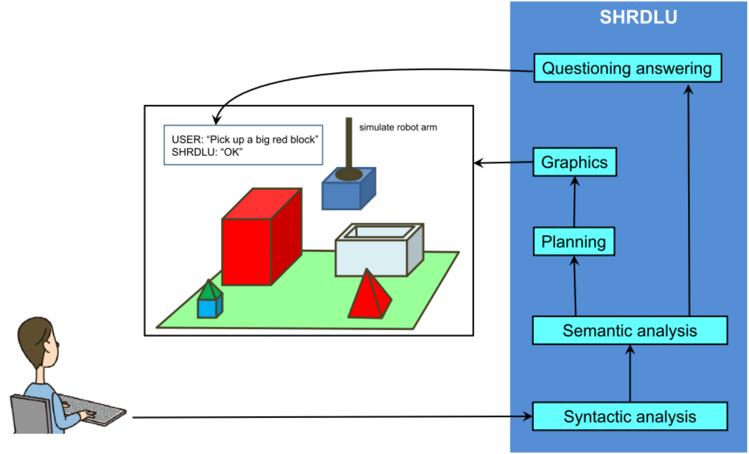
BASEBALL(1961) and LUNAR(1971-1973)

- Two early question answering systems were BASEBALL and LUNAR.
- BASEBALL answered questions about the US baseball league over a period of one year. LUNAR, in turn, answered questions about the geological analysis of rocks returned by the Apollo moon missions.
- Both question answering systems were very effective in their chosen domains. In fact, LUNAR was demonstrated at a lunar science convention in 1971 and it was able to answer 90% of the questions in its domain posed by people untrained on the system.

SHRDLU (1968-1970)

- SHRDLU was an early natural language understanding computer program, developed by Terry Winograd at MIT in 1968–1970.
- In it, the user carries on a conversation with the computer, moving objects, naming collections and querying the state of a simplified "blocks world", essentially a virtual box filled with different blocks.
- SHRDLU was a famous system in the AI history, which firstly demonstrated the ability of an AI system to interact with humans with natural languages to undertake certain tasks.

SHRDLU (1968-1970)



Winograd, Terry, *Understanding Natural Language*, New York: Academic Press, 1972. http://hci.stanford.edu/~winograd/shrdlu/

Eliza (1964)

- ELIZA is an early natural language processing computer program created from 1964 to 1966 at the MIT Artificial Intelligence Laboratory by Joseph Weizenbaum.
- Eliza simulated conversation by using a "pattern matching" and substitution methodology that gave users an illusion of understanding on the part of the program, but had no built in framework for contextualizing events.
- Many early users were convinced of ELIZA's intelligence and understanding, despite Weizenbaum's insistence to the contrary.
- ELIZA was one of the first chatterbots and one of the first programs capable of attempting the Turing test.

Eliza (1964)

```
Welcome to
                    EEEEEE  LL        IIII  ZZZZZZZ   AAAAA
                    EE      LL         II       ZZ   AA   AA
                    EEEEE   LL         II     ZZZ    AAAAAAA
                    EE      LL         II    ZZ      AA   AA
                    EEEEEE  LLLLLL    IIII  ZZZZZZZ  AA   AA

Eliza is a mock Rogerian psychotherapist.
The original program was described by Joseph Weizenbaum in 1966.
This implementation by Norbert Landsteiner 2005.


ELIZA: Is something troubling you ?
YOU:   Men are all alike.
ELIZA: What is the connection, do you suppose ?
YOU:   They're always bugging us about something or other.
ELIZA: Can you think of a specific example ?
YOU:   Well, my boyfriend made me come here.
ELIZA: Is it important to you that your boyfriend made you come here ?
YOU:   He says I'm depressed much of the time.
ELIZA: I am sorry to hear that you are depressed.
YOU:   It's true. I am unhappy.
ELIZA: Can you explain what made you unhappy ?
YOU:   
```

Expert Systems (1970s-1980s)

- An expert system is a computer system that emulates the decision-making ability of a human expert.
- Expert systems solve complex problems by reasoning through bodies of knowledge, represented mainly as if–then rules.
- Expert systems rely heavily on expert-constructed and organized knowledge bases.
- Expert systems were firstly created in the 1970s and then proliferated in the 1980s.

Content

Big data era: open domain QA

- Open Domain QA and Web-based QA
- Community QA (2000s)
- Wolfram Alpha (2009)
- IBM Watson (2011)

Open domain QA

- 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: a brief history

- Simmons et al. (1964) did first exploration of answering questions from an expository text based on matching dependency parses of a question and answer
- Murax (Kupiec 1993) aimed to answer questions over an online encyclopedia using IR and shallow linguistic processing
- The NIST TREC QA track begun in 1999 first rigorously investigated answering fact questions over a large collection of documents
- IBM's Jeopardy! System (DeepQA, 2011) brought attention to a version of the problem; it used an ensemble of many methods
- DrQA (Chen et al. 2016) uses IR followed by neural reading comprehension to bring deep learning to Open-domain QA

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

Web-based QA

- As a special case of open-domain QA, big search engines like Google provide direct answers to user queries rather than a list of web pages, when it feels confident.

Web-based QA

how many countries are adjacent to china in land?

All | Images | Maps | News | Videos | More | Settings | Tools

About 638,000,000 results (0.72 seconds)

14

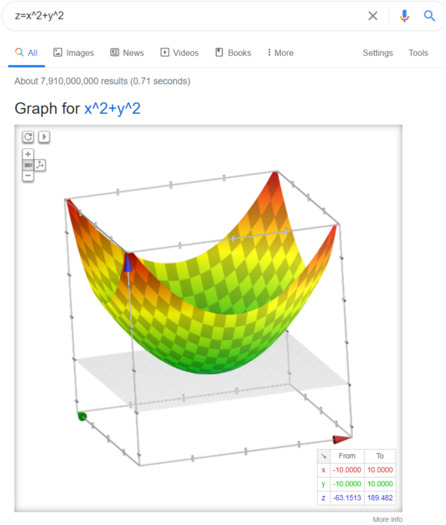
As the most populous **country** in the world and third largest in area, **China** also has the largest number of neighbours (14) sharing its 22,000km **land** borders namely: North Korea, Russia, Mongolia, Kazakhstan, Kyrgyzstan, Tajikistan, Afghanistan, Pakistan, India, Nepal, Bhutan, Myanmar, Laos and Vietnam. Mar 1, 2012

www.eu-asiacentre.eu › pub_details

China and its Neighbours: troubled relations - eu asia centre

About Featured Snippets | Feedback

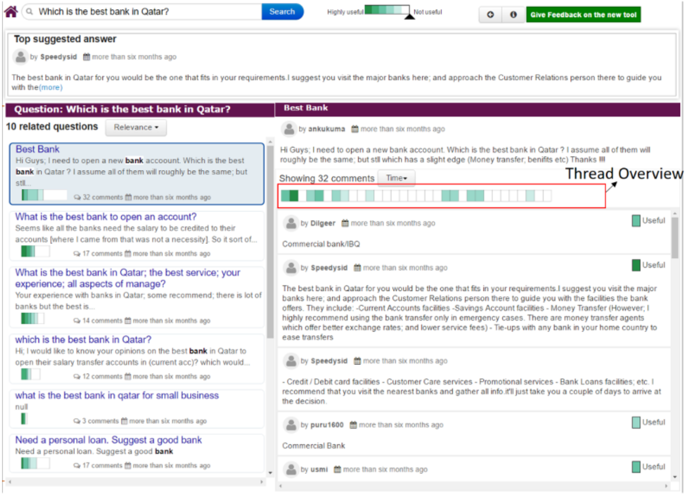
Web-based QA


z=x^2+y^2
All
Images
News
Videos
Books
More
Settings
Tools
About 7,910,000,000 results (0.71 seconds)
Graph for x^2+y^2
From
To
x -10.0000 10.0000
y -10.0000 10.0000
z -63.1513 189.482
More info

Community QA

- Community QA is a category of QA which is based on QA forums.
- A community QA system searches the forum to find an existing quesiton which is equivalent to the user's input and return a best answer to that question from the forum.
- The questions which can be answered by a community QA system are limited to the forum data, however, the answers are generally of high quality because it is written by humans.
- Community Question Answering has seen a spectacular increase in popularity in 2000s along with the popularity of QA forums like Yahoo! Answers, Stack Overflow, Quora, etc.

Community QA

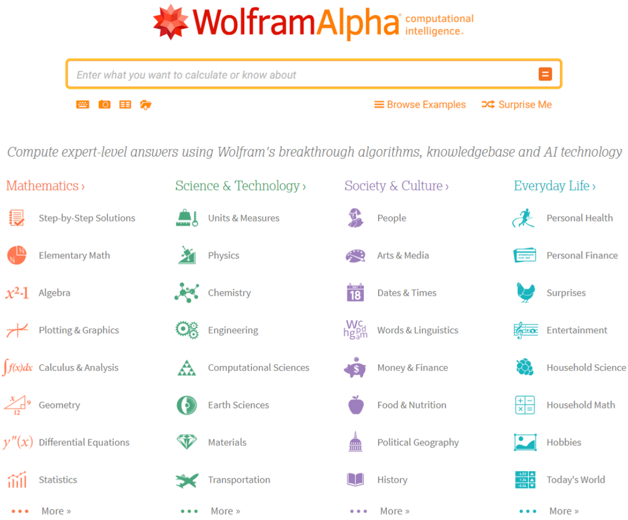


Community Question Answering System, NTU NLP Group

Wolfram Alpha (2009)

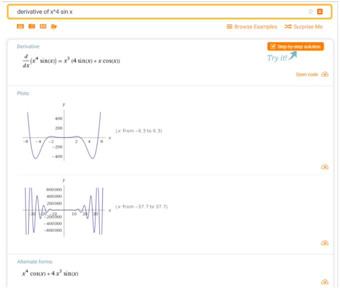
- WolframAlpha (also styled Wolfram|Alpha) is a computational knowledge engine or answer engine developed by Wolfram Alpha LLC, a subsidiary of Wolfram Research.
- It is an online service that answers factual queries directly by computing the answer from externally sourced "curated data", rather than providing a list of documents or web pages that might contain the answer as a search engine might.
- WolframAlpha can only provide robust query results based on computational facts, not queries on the social sciences, cultural studies or even many questions about history where responses require more subtlety and complexity.

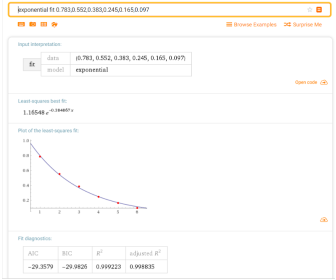
Wolfram Alpha (2009)



Wolfram Alpha (2009)

- WolframAlpha can not only answer factual questions with texts, but also run programs to compute the answer and give visualization of the answers (e.g. a curve of a function).





IBM Watson (2011)

- Watson is a question-answering computer system capable of answering questions posed in natural language, developed in IBM's DeepQA project by a research team led by principal investigator David Ferrucci.
- In 2011, the Watson computer system competed on Jeopardy! against champions Brad Rutter and Ken Jennings, winning the first place prize of $1 million.
- In February 2013, IBM announced that Watson software system's first commercial application would be for utilization management decisions in lung cancer treatment at Memorial Sloan Kettering Cancer Center, New York City, in conjunction with WellPoint (now Anthem).

IBM Watson in Jeopardy! (2013)

This is another breakthrough of an AI system to beat human competitors in real games after IBM's DeepBlue defeated the world chess champion Garry Kasparov on 10 February 1996.

Content

Neural era: machine comprehension and dialog systems

- Machine reading comprehension
- Siri(2011) and voice assistants
- Amazon Alexa(2015) and smart speakers
- XiaoIce(2014) and social chatbots
- Natural Language Interface to Databases
- Knowledge Base/Graph QA (KBQA/KGQA)
- Visual QA / Visual Dialog

Machine reading comprehension

- 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.

Machine reading comprehension

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

Rajpurkar, Pranav, et al. "Squad: 100,000+ questions for machine comprehension of text." arXiv:1606.05250, 2016

Machine reading comprehension: a brief history

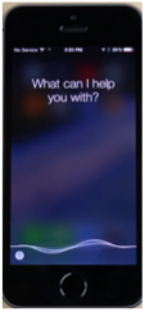
- Much early NLP work attempted reading comprehension
  - Schank, Abelson, Lehnert et al. c. 1977 – "Yale A.I. Project"
- Revived by Lynette Hirschman in 1999:
  - Could NLP systems answer human reading comprehension questions for 3rd to 6th graders? Simple methods attempted.
- Revived again by Chris Burges in 2013 with MCTest
  - Again answering questions over simple story texts
- Floodgates opened in 2015/16 with the production of large datasets which permit supervised neural systems to be built
  - Hermann et al. (NIPS 2015) DeepMind CNN/DM dataset
  - Rajpurkar et al. (EMNLP 2016) SQuAD
  - MS MARCO, TriviaQA, RACE, NewsQA, NarrativeQA, ...

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

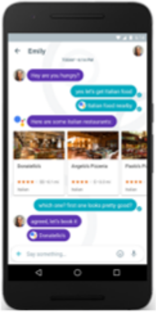
Siri (2011) and voice assistants

- Siri is a virtual assistant that is part of Apple Inc.'s operating systems.
- The assistant uses voice queries and a natural-language user interface to answer questions, make recommendations, and perform actions by delegating requests to a set of Internet services.
- Siri became the first digital virtual assistant to be standard on a smartphone when the iPhone 4s came out on October 4, 2011, and followed by a number of similar products including Google Assistant and Microsoft Cortona.

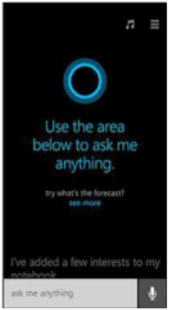
Siri (2011) and voice assistants



Apple
Siri
2011



Google
Assistant
2016



Microsoft
Cortana
2014

Amazon Alexa (2015) and smart speakers

- Amazon Alexa, also known simply as Alexa, is a virtual assistant AI technology developed by Amazon, first used in the Amazon Echo smart speakers developed by Amazon Lab126.
- It is capable of voice interaction, music playback, making to-dolists, and many other real-time information. Alexa can also control several smart devices using itself as a home automation system.
- Users are able to extend the Alexa capabilities by installing "skills" (additional functionality developed by third-party vendors, in other settings more commonly called apps such as weather programs and audio features).

Amazon Alexa (2015) and smart speakers

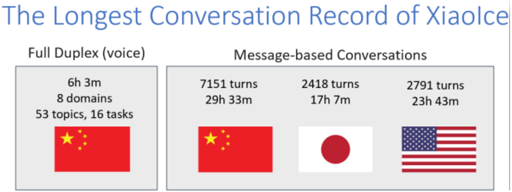
- Amazon was successful in the market and has been followed by other smart speaker products like Google Home, etc.

Microsoft XiaoIce (2014) and social chatbots

- Xiaoice (Chinese: 微软小冰) is the AI system developed by Microsoft STCA in 2014 based on emotional computing framework.
- Through the comprehensive application of algorithms, cloud computing and big data, Xiaoice adopts the intergenerational upgrade method to gradually form a complete artificial intelligence system to EQ.
- Microsoft Xiaoice has become one of the world's biggest interdisciplinary AI systems and taken various product forms including chat bot, intelligent voice assistant, AI content creation and production platform, etc.

Microsoft XiaoIce (2014) and social chatbots

- In many countries around the world, as a single brand, Microsoft Xiaoice has covered 660 million on-line users, 450 million third-party IoT devices and 900 million content viewers. It has kept 23 conversations per session (CPS) averagely with users while greatly increasing interactive scenarios.



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

Natural Language Interface to Databases

- Convert a natural language question to a database query
- Execute the database query on the database and obtain the answer
- Example systems:
  - Luner (1960s) NLMENU(1980s)
  - PRECISE (2002) ThoughtSpot (2012) Arimo (2012)
  - NaLIR(2014) Poser BI (2015) SimpleQL (2013)

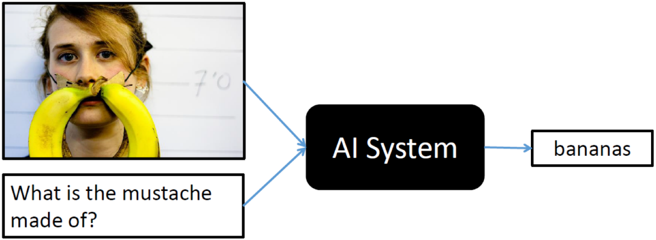
Jonas Chapuis, Natural Language Interfaces to Databases (NLIDB)

Knowledge Base/Graph QA (KBQA/KGQA)

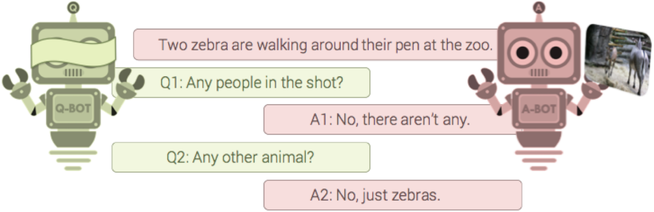
- QA over knowledge graphs
  - Freebase
  - DBPedia
  - etc.
- Large scale data
- Structured data
- Inference is needed for answering some questions

Visual QA / Visual Dialog

Visual QA



Visual Dialog



Visual Question Answering and Dialog Workshop

Content

Why we care about QA?

Because QA is awesome

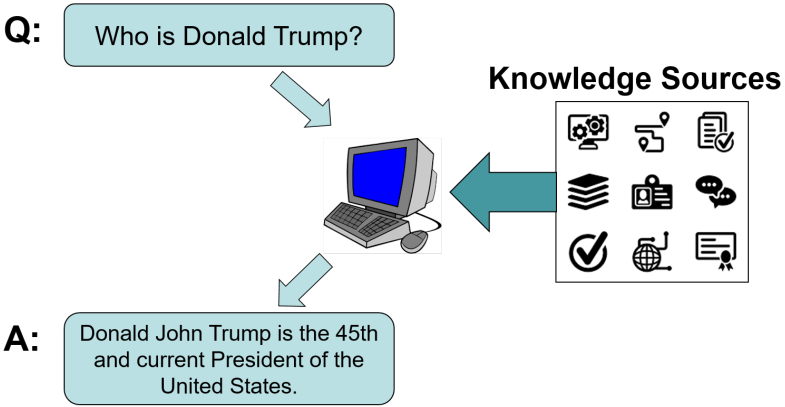
1. QA is an AI-complete problem.
   If we solve QA, we have solved every other problem, too.
2. Many immediate and obvious applications
   Search, dialogue, information extraction, summarisation, ...
3. Some pretty nice results already
   IBM Watson and Jeopardy!, Siri, Google Search ...
4. Lots left to do!
   Plenty of interesting research and hard problems as well as low-hanging fruit.

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

Content

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

Question answering



Content

3 Question Answering

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

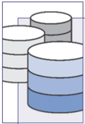
Knowledge sources

Text

Rules & Facts

Web

QA Archive



Database

Knowledge Graph

Rules and Facts

- Rules:
  - R1: $\mathrm{Man}(x) \rightarrow \mathrm{Mortal}(x)$
  - R2: $\mathrm{Bird}(x) \rightarrow \mathrm{CanFly}(x)$
- Facts:
  - F1: $\mathrm{Man}(Socrates)$
  - F2: $\mathrm{Man}(Plato)$
  - F3: $\mathrm{Bird}(Parrot_A)$
- Typically used in expert systems.

QA Archive

- FAQ list
- Question Answering forums:
  - Yahoo! Answers
  - Stack Overflow
  - Quora
  - ...

Stack Overflow - Questions

Java - How can I disable a TLS cipher for only some protocols using JVM Config?

Ask Question

10

4

I've seen lots of examples of disabling TLS ciphers in java using jdk.tls.disabledAlgorithms, for example:

```
jdk.tls.disabledAlgorithms=MD2, RSA keySize < 1024, TLS_ECDHE_RSA_WITH_AES_128_CBC_SHA256
```

But how can I disable a cipher for only certain protocols, using `jdk.tls.disabledAlgorithms` or a similar config?

For example, how can I disable `TLS_ECDHE_RSA_WITH_AES_128_CBC_SHA256` for `TLSv1.1` only?

It doesn't seem to support the `openssl` way of doing this, which is like so:

```
TLSv1.1:!TLS_ECDHE_RSA_WITH_AES_128_CBC_SHA256
```

It doesnt cause any errors, but the cipher is still allowed.

EDIT: Note that I'm only really interested in JVM config based answers, as I don't control the code that's on lots of these servers, just the JVM and JVM configurations. Some are even 3rd party servers, so more of an ops level thing than anything.

EDIT 2: Note that you can run a java app and supply arguments that change which protocols and ciphers are used, e.g. `java -server -Djava.security.properties=./my/custom/java.security -jar myapp.jar` will do it - but it wont let you filter ciphers by protocol, only ciphers, or protocols, from what I can see. The file would contain a property entry like `jdk.tls.disabledAlgorithms`

java java-8

asked 7 days ago
viewed 149 times
active today

Stack Overflow - Answers

2 Answers

active oldest votes

2

JSSE docs say that the `https.protocols` property can store comma separated list of supported protocols in a given SSL context, however this property is used by current JSSE implementation, but could be disregarded by other vendors or future versions, so YMMV.

Programatically you can achieve it like so:

```
SSLSocket socket = (SSLSocket) SSLSocketFactory.getDefault().createSocket();
socket.setEnabledCipherSuites(new String[] {
    CipherSuite.TLS_RSA_WITH_RC4_128_MD5.javaName,
    CipherSuite.TLS_RSA_WITH_RC4_128_SHA.javaName,
});

//allow TLS1.2 only
socket.setEnabledProtocols(new String[] {
    TlsVersion.TLS_1_2.javaName,
});
```

share improve this answer

edited Oct 15 at 10:53

answered Oct 15 at 10:46

diginoise

4,169 • 1 • 15 • 26

1 Thanks for the response... I'm mostly interested in config only options though, as I don't control the code that's on lots of these servers, just the JVM and JVM configurations. Some are even 3rd party servers, so more of an ops thing than anything. If I don't get any config based answers, I'll give it to you ;-) I've added this as a clarification to my question. Thanks! – Brad Parks Oct 16 at 10:43

add a comment

Knowledge Graph

*"The **Knowledge Graph** is a knowledge base used by Google to enhance its search engine's search results with semantic-search information gathered from a wide variety of sources."*

"A Knowledge graph (i) mainly describes real world entities and interrelations, organized in a graph (ii) defines possible classes and relations of entities in a schema" (iii) allows potentially interrelating arbitrary entities with each other... [Paulheim H.]

"We defines a Knowledge Graph as an RDF graph consists of a set of RDF triples where each RDF triple (s,p,o) is an ordered set of following RDF term ...." *[Pujara J. al al.]*

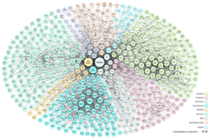
Nitish Aggarwal, et al., Knowledge Graphs: In Theory and Practice, CIKM17 Tutorials

Knowledge Graph

LinkedIn Knowledge Graph

Freebase

DBpedia



yago select knowledge

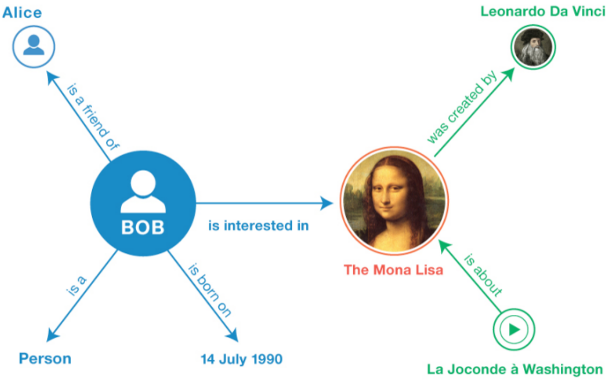
Amazon Product Graph

Facebook Entity Graph

Microsoft Satori

Nitish Aggarwal, et al., Knowledge Graphs: In Theory and Practice, CIKM17 Tutorials

A (very small) Knowledge Graph



http://www.w3.org/TR/2014/NOTE-rdf11-primer-20140225/example-graph.jpg

DBpedia

- The English version of the DBpedia knowledge base describes 4.58 million things, out of which 4.22 million are classified in a consistent ontology, including 1,445,000 persons, 735,000 places (including 478,000 populated places), 411,000 creative works (including 123,000 music albums, 87,000 films and 19,000 video games), 241,000 organizations (including 58,000 companies and 49,000 educational institutions), 251,000 species and 6,000 diseases.
- In addition, we provide localized versions of DBpedia in 125 languages. All these versions together describe 38.3 million things, out of which 23.8 million are localized descriptions of things that also exist in the English version of DBpedia.

DBpedia – Linked Data Fragments

DBpedia 2016-04

Query DBpedia 2016-04 by triple pattern

subject:

predicate:

object: "1899-05-06"^^http://www.w3.org/2001/XMLSchema#date

Find matching triples

Matches in DBpedia 2016-04 for *{ ?s ?p "1899-05-06"^^http://www.w3.org/2001/XMLSchema#da...*

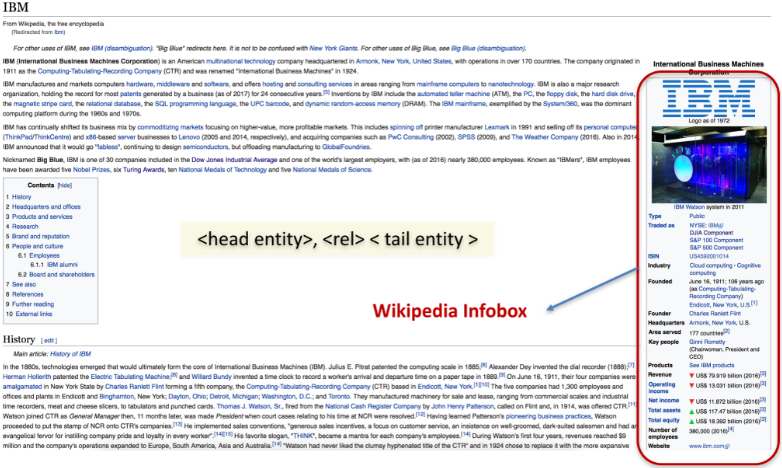
Showing triples 1 to 74 of 74 with 100 triples per page.

```
Billy_Cotton  birthDate  "1899-05-06".
Charlie_Irvis  birthDate  "1899-05-06".
Edward_Grahame_Johnstone  birthDate  "1899-05-06".
Eiliv_Austlid  birthDate  "1899-05-06".
Julio_Vega_Batlle  birthDate  "1899-05-06".
Kenneth_Hand  birthDate  "1899-05-06".
Ralyn_M._Hill  birthDate  "1899-05-06".
Tommy_Magee  birthDate  "1899-05-06".
Wally_Sieb  birthDate  "1899-05-06".
Alfred_Pillinger  deathDate  "1899-05-06".
Earley_F._Poppleton  deathDate  "1899-05-06".
Edward_Butterfield  deathDate  "1899-05-06".
```

DBpedia: Knowledge extraction

Nitish Aggarwal, et al., Knowledge Graphs: In Theory and Practice, CIKM17 Tutorials

DBpedia: Knowledge extraction


IBM

From Wikipedia, the free encyclopedia
(Redirected from Ibm)

For other uses of IBM, see IBM (disambiguation). "Big Blue" redirects here. It is not to be confused with New York Giants. For other uses of Big Blue, see Big Blue (disambiguation).

IBM (**International Business Machines Corporation**) is an American multinational technology company headquartered in Armonk, New York, United States, with operations in over 170 countries. The company originated in 1911 as the Computing-Tabulating-Recording Company (CTR) and was renamed "International Business Machines" in 1924.

IBM manufactures and markets computers hardware, middleware and software, and offers hosting and consulting services in areas ranging from mainframe computers to nanotechnology. IBM is also a major research organization, holding the record for most patents generated by a business (as of 2017) for 24 consecutive years.[5] Inventions by IBM include the automated teller machine (ATM), the PC, the floppy disk, the hard disk drive, the magnetic stripe card, the relational database, the SQL programming language, the UPC barcode, and dynamic random-access memory (DRAM). The IBM mainframe, exemplified by the System/360, was the dominant computing platform during the 1960s and 1970s.

IBM has continually shifted its business mix by commoditizing markets focusing on higher-value, more profitable markets. This includes spinning off printer manufacturer Lexmark in 1991 and selling off its personal compute (ThinkPad/ThinkCentre) and x86-based server businesses to Lenovo (2005 and 2014, respectively), and acquiring companies such as PwC Consulting (2002), SPSS (2009), and The Weather Company (2016). Also in 201 IBM announced that it would go "fabless", continuing to design semiconductors, but offloading manufacturing to GlobalFoundries.

Nicknamed **Big Blue**, IBM is one of 30 companies included in the Dow Jones Industrial Average and one of the world's largest employers, with (as of 2016) nearly 380,000 employees. Known as "IBMers", IBM employees have been awarded five Nobel Prizes, six Turing Awards, ten National Medals of Technology and five National Medals of Science.

Contents [hide]

History [edit]

Main article: History of IBM

In the 1880s, technologies emerged that would ultimately form the core of International Business Machines (IBM). Julius E. Pitrat patented the computing scale in 1885;[6] Alexander Dey invented the dial recorder (1888);[7] Herman Hollerith patented the Electric Tabulating Machine;[8] and Willard Bundy invented a time clock to record a worker's arrival and departure time on a paper tape in 1889.[9] On June 16, 1911, their four companies were amalgamated in New York State by Charles Ranlett Flint forming a fifth company, the Computing-Tabulating-Recording Company (CTR) based in Endicott, New York.[1][10] The five companies had 1,300 employees and offices and plants in Endicott and Binghamton, New York; Dayton, Ohio; Detroit, Michigan; Washington, D.C.; and Toronto. They manufactured machinery for sale and lease, ranging from commercial scales and industrial time recorders, meat and cheese slicers, to tabulators and punched cards. Thomas J. Watson, Sr., fired from the National Cash Register Company by John Henry Patterson, called on Flint and, in 1914, was offered CTR.[11] Watson joined CTR as *General Manager* then, 11 months later, was made *President* when court cases relating to his time at NCR were resolved.[12] Having learned Patterson's pioneering business practices, Watson proceeded to put the stamp of NCR onto CTR's companies.[13] He implemented sales conventions, "generous sales incentives, a focus on customer service, an insistence on well-groomed, dark-suited salesmen and had a evangelical fervor for instilling company pride and loyalty in every worker".[14][15] His favorite slogan, "THINK", became a mantra for each company's employees.[14] During Watson's first four years, revenues reached $9 million and the company's operations expanded to Europe, South America, Asia and Australia.[14] "Watson had never liked the clumsy hyphenated title of the CTR" and in 1924 chose to replace it with the more expansive

International Business Machines Corporation

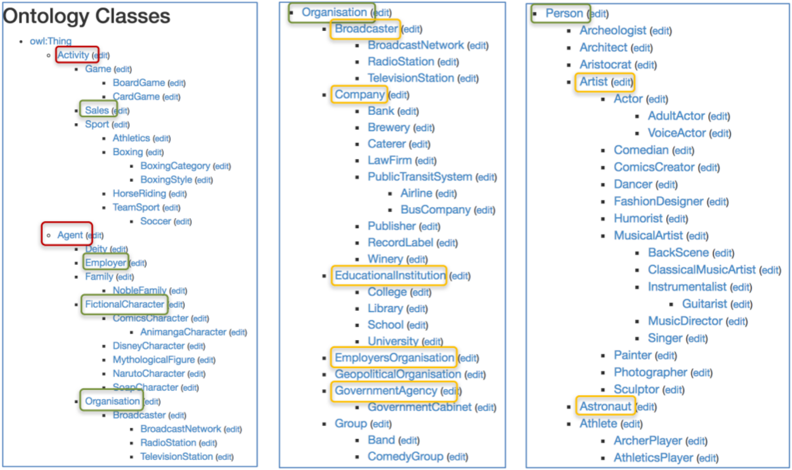
Logo as of 1972

IBM Watson system in 2011

Type	Public
Traded as	NYSE: IBM DJIA Component S&P 100 Component S&P 500 Component
ISIN	US4592001014
Industry	Cloud computing · Cognitive computing
Founded	June 16, 1911; 106 years ago (as Computing-Tabulating-Recording Company) Endicott, New York, U.S.[1]
Founder	Charles Ranlett Flint
Headquarters	Armonk, New York, U.S.
Area served	177 countries[2]
Key people	Ginni Rometty (Chairwoman, President and CEO)
Products	See IBM products
Revenue	▼ US$ 79.919 billion (2016)[3]
Operating income	▼ US$ 13.031 billion (2016)[3]
Net income	▼ US$ 11.872 billion (2016)[3]
Total assets	▲ US$ 117.47 billion (2016)[3]
Total equity	▲ US$ 18.392 billion (2016)[3]
Number of employees	380,000 (2016)[4]
Website	www.ibm.com

Nitish Aggarwal, et al., Knowledge Graphs: In Theory and Practice, CIKM17 Tutorials

DBpedia: Knowledge extraction



Ontology Classes

- owl:Thing
  - Activity (edit)
    - Game (edit)
      - BoardGame (edit)
      - CardGame (edit)
    - Sales (edit)
    - Sport (edit)
      - Athletics (edit)
      - Boxing (edit)
        - BoxingCategory (edit)
        - BoxingStyle (edit)
      - HorseRiding (edit)
      - TeamSport (edit)
        - Soccer (edit)
  - Agent (edit)
    - Deity (edit)
    - Employer (edit)
    - Family (edit)
      - NobleFamily (edit)
    - FictionalCharacter (edit)
      - ComicsCharacter (edit)
        - AnimangaCharacter (edit)
      - DisneyCharacter (edit)
      - MythologicalFigure (edit)
      - NarutoCharacter (edit)
      - SoapCharacter (edit)
    - Organisation (edit)
      - Broadcaster (edit)
        - BroadcastNetwork (edit)
        - RadioStation (edit)
        - TelevisionStation (edit)

- Organisation (edit)
  - Broadcaster (edit)
    - BroadcastNetwork (edit)
    - RadioStation (edit)
    - TelevisionStation (edit)
  - Company (edit)
    - Bank (edit)
    - Brewery (edit)
    - Caterer (edit)
    - LawFirm (edit)
    - PublicTransitSystem (edit)
      - Airline (edit)
      - BusCompany (edit)
    - Publisher (edit)
    - RecordLabel (edit)
    - Winery (edit)
  - EducationalInstitution (edit)
    - College (edit)
    - Library (edit)
    - School (edit)
    - University (edit)
  - EmployersOrganisation (edit)
  - GeopoliticalOrganisation (edit)
  - GovernmentAgency (edit)
    - GovernmentCabinet (edit)
  - Group (edit)
    - Band (edit)
    - ComedyGroup (edit)

- Person (edit)
  - Archeologist (edit)
  - Architect (edit)
  - Aristocrat (edit)
  - Artist (edit)
    - Actor (edit)
      - AdultActor (edit)
      - VoiceActor (edit)
    - Comedian (edit)
    - ComicsCreator (edit)
    - Dancer (edit)
    - FashionDesigner (edit)
    - Humorist (edit)
    - MusicalArtist (edit)
      - BackScene (edit)
      - ClassicalMusicArtist (edit)
      - Instrumentalist (edit)
        - Guitarist (edit)
      - MusicDirector (edit)
      - Singer (edit)
    - Painter (edit)
    - Photographer (edit)
    - Sculptor (edit)
  - Astronaut (edit)
  - Athlete (edit)
    - ArcherPlayer (edit)
    - AthleticsPlayer (edit)

Nitish Aggarwal, et al., Knowledge Graphs: In Theory and Practice, CIKM17 Tutorials

DBpedia: Knowledge extraction

Properties on *Actor*:

Name	Label	Domain	Range
academyAward (edit)	Academy Award	Artist	Award
afiAward (edit)	AFI Award	Artist	Award
arielAward (edit)	Ariel Award	Actor	Award
associatedAct (edit)	associated act	Artist	Artist
baftaAward (edit)	BAFTA Award	Artist	Award
cesarAward (edit)	Cesar Award	Artist	Award
disciple (edit)	disciple	Artist	Artist
dutchRKDCode (edit)	Dutch RKD code	Artist	*xsd:string*
emmyAward (edit)	Emmy Award	Artist	Award
field (edit)	field	Artist	owl:Thing
filmFareAward (edit)	Film Fare Award	Artist	Award
gaudiAward (edit)	Gaudí Award	Artist	Award
geminiAward (edit)	Gemini Award	Actor	Award
goldenCalfAward (edit)	Golden Calf Award	Actor	Award
goldenGlobeAward (edit)	Golden Globe Award	Artist	Award
goldenRaspberryAward (edit)	Golden Raspberry Award	Actor	Award
goyaAward (edit)	Goya Award	Artist	Award
grammyAward (edit)	Grammy Award	Artist	Award
iftaAward (edit)	IFTA Award	Actor	Award

Nitish Aggarwal, et al., Knowledge Graphs: In Theory and Practice, CIKM17 Tutorials

Content

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

Knowledge representation

=> Linguistic Annotations

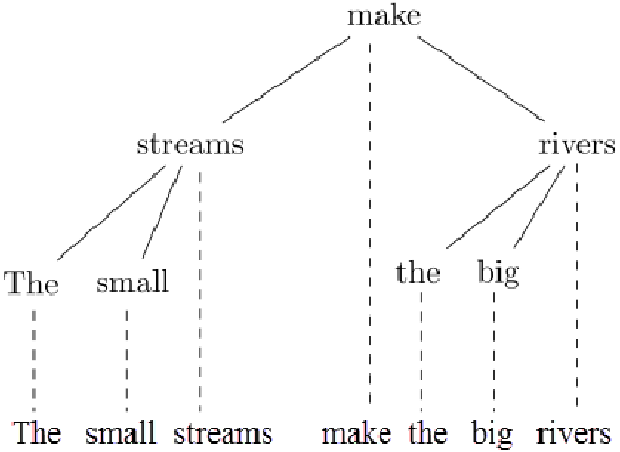
=> Statistical Representation

=> Neural Representation

Linguistic annotations

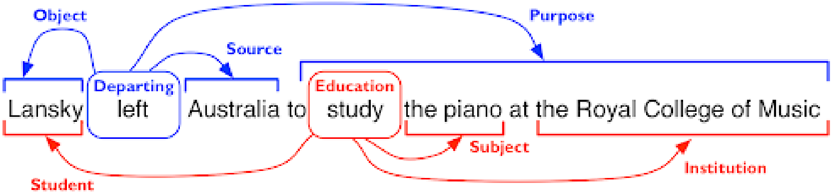
- Part-of-Speech Tagging
- Syntactic Parsing: Constituent or Dependency
- Named Entities
- Entity Relations
- Events
- Anaphora, Co-Reference and Mentions
- Rhetorical Structures
- Semantic Role Labelling
- Word Senses
- Sentiments

Dependency trees



Haitao Liu, et al. "Dependency distance: a new perspective on syntactic patterns in natural languages." Physics of life reviews 21 2017

Semantic role labeling



http://ivan-titov.org/topics/semantics.html

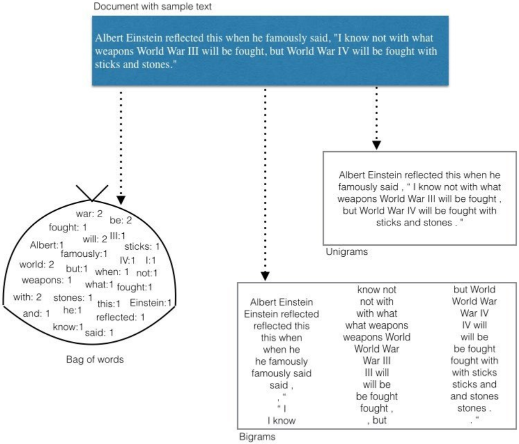
Knowledge representation

=> Linguistic Annotations

=> Statistical Representation

=> Neural Representation

Bag-of-Words and Bag-of-Ngrams



https://qph.fs.quoracdn.net/main-qimg-c47060d2f02439a44795e2fbcf2ca347.webp

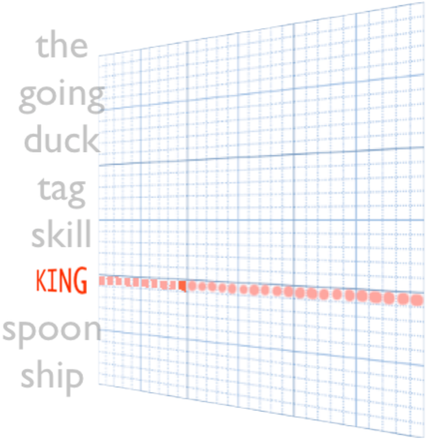
Vector Space Model and TF-IDF

- Vector space model represent a document as a high dimensional vector rather than a set (bag).
- TF-IDF is the most commonly used weighing schema in VSM in information retrieval
  - TF: term frequency
  - IDF: inverse document frequence

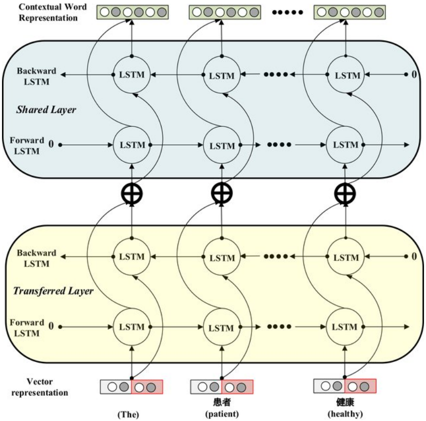
Knowledge representation

- => Linguistic Annotations
- => Statistical Representation
- => Neural Representation

Word embeddings



Contextual word embeddings



Dong et al., Deep learning for named entity recognition on Chinese electronic medical records: Combining deep transfer learning with multitask bi-directional LSTM RNN, PLoS ONE 14(5):e0216046, 2019

Sentence embeddings

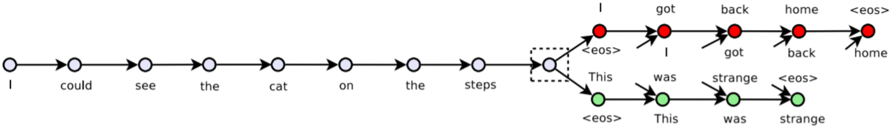
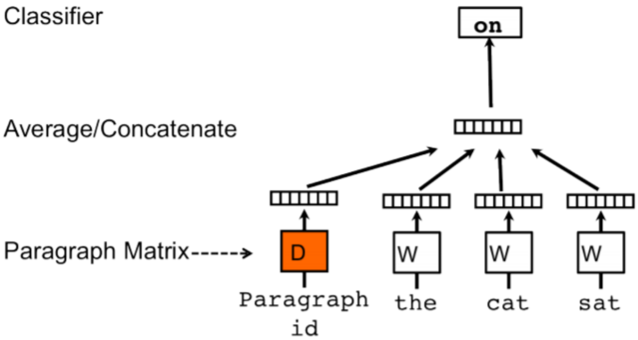


Figure 1: The skip-thoughts model. Given a tuple (s_{i-1}, s_i, s_{i+1}) of contiguous sentences, with s_i the i-th sentence of a book, the sentence s_i is encoded and tries to reconstruct the previous sentence s_{i-1} and next sentence s_{i+1}. In this example, the input is the sentence triplet *I got back home. I could see the cat on the steps. This was strange.* Unattached arrows are connected to the encoder output. Colors indicate which components share parameters. $\langle \text{eos} \rangle$ is the end of sentence token.

Kiros, Ryan, et al. "Skip-thought vectors." Advances in neural information processing systems. 2015

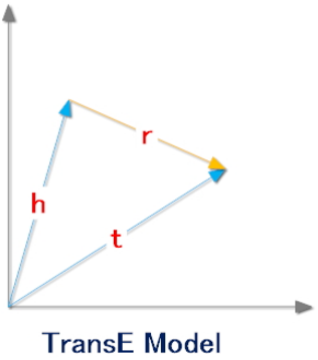
Document embeddings

Doc2Vec



A gentle introduction to Doc2Vec

Knowledge embeddings



Bordes, et al., Translating embeddings for modeling multi-relational data, Adv. Neural Inf. Process. Syst., 2013

Content

Question understanding

- Pattern Matching
- Similar Question Retrieval
- Question Paraphrasing
- Question Classification and Slot Filling
- Semantic Parsing

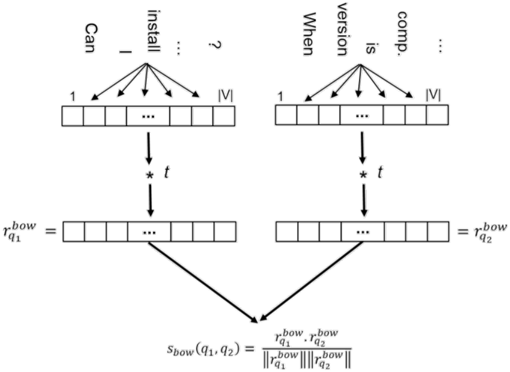
Pattern matching

- Example: Who is X of Y?
  - Who is the first president of US?
  - Who is the second husband of Wendi Deng?
- Accurate Understanding
- Low Coverage
- Only suitable for very small domain

Equivalent Question Retrieval

- Basic ideas:
  - Retrieve existing questions with the query using an IR technologies;
  - Select the best one from the top ranked question returned by the IR system, using some similarity metrics;
- Can be used in scenarios like community QA or open-domain QA.

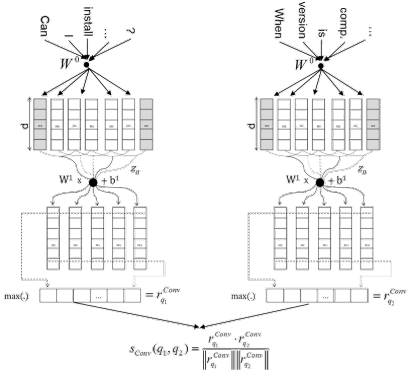
Equivalent Question Retrieval



Question retrieval with bag-of-words similarity

Dos Santos, et al., Learning hybrid representations to retrieve semantically equivalent questions. ACL-IJCNLP2015

Equivalent Question Retrieval

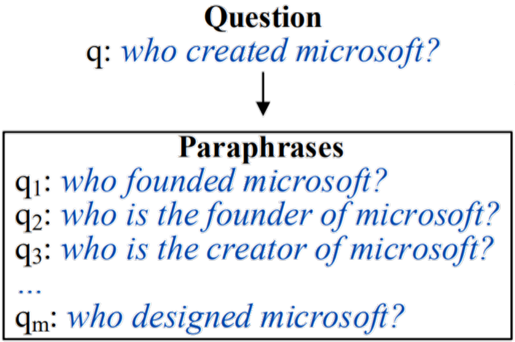


Question retrieval with convolutional similarity

Dos Santos, et al., Learning hybrid representations to retrieve semantically equivalent questions. ACL-IJCNLP2015

Question paraphrasing

Question Paraphrasing is effective for most QA tasks.



Li Dong, et al„ Learning to Paraphrase for Question Answering, arXiv:1708.06022, 2017

Question classification

- Classification by answer types:
  - Factual questions: What is the largest city ... ?
  - Opinions: What is the authors' attitude ...?
  - Summaries: What are the arguments for and against ...?
- Classification by speech acts:
  - Yes/No questions: Is it true that ...?
  - WH questions: Who was the first president ...?
  - Indirect request: I would like you to list ... .
  - Commands: Name all presidents ... .
- Complex / difficult questions:
  - Why/How questions: How to build a house?
  - What questions: What did they do?

Question classification

Class	#	Class	#
ABBREV.	9	description	7
abb	1	manner	2
exp	8	reason	6
ENTITY	94	**HUMAN**	65
animal	16	group	6
body	2	individual	55
color	10	title	1
creative	0	description	3
currency	6	**LOCATION**	81
dis.med.	2	city	18
event	2	country	3
food	4	mountain	3
instrument	1	other	50
lang	2	state	7
letter	0	**NUMERIC**	113
other	12	code	0
plant	5	count	9
product	4	date	47
religion	0	distance	16
sport	1	money	3
substance	15	order	0
symbol	0	other	12
technique	1	period	8
term	7	percent	3
vehicle	4	speed	6
word	0	temp	5
DESCRIPTION	138	size	0
definition	123	weight	4

The distribution of 500 TREC 10 questions over the question hierarchy.

Xin Li and Dan Roth. "Learning question classifiers." COLING 2002.

Slot filling

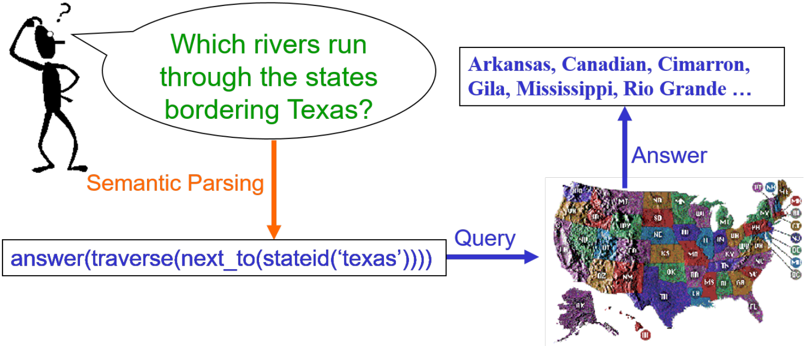
- Question classification gives a fine grained understanding of user's intent.
- Give a class of questions, there are slots to be filled to give the details of the question.
- Example:
  - Input: When did the Second World War break out?
  - Output:
    - Classification: TIME
    - Slots:
      Events: BREAK_OUT
      Subjects: the Second World War

Semantic parsing

- Convert a natural language sentence to a semantic representation, for example:
  - Logic forms
  - Lambda Calculus
  - SQL
  - SPARQL
  - Programming language scripts: Prolog, Lisp, Python, etc.

Semantic parsing: An example

- Query application for U.S. geography database containing about 800 facts [Zelle & Mooney, 1996]



Raymond J. Mooney, Semantic Parsing for Question Answering (slides).

Content

Answer generation

- Answer retrieval
- Answer extraction
- Answer ranking
- Database querying
- Knowledgebase query
- Rule-based inference engine
- Inference over knowledge graphs
- Answer Generation

Answer retrieval

- Find the paragraphs which are related to the question, from a large collection text
- Approaches:
  - Build a paragraph-based search engine.
  - Use an existing search engine to find a list of documents, then find the most relevant paragraphs
- A subtask for Web-based QA (Open domain QA)
- Technologies: IR

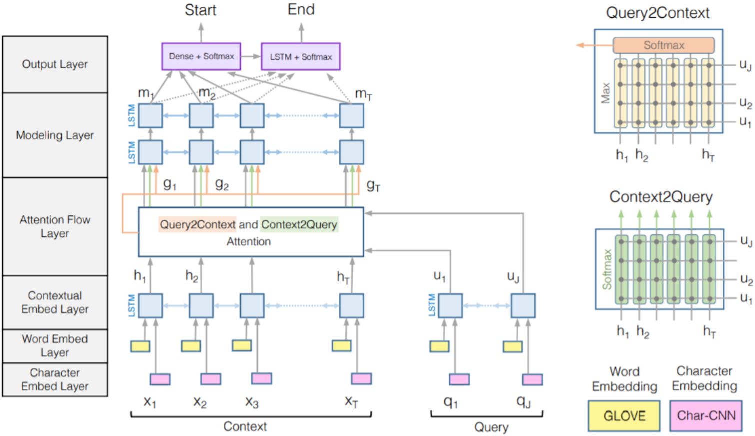
Answer ranking

- When multiple answer candidates are obtained / generated, a ranking algorithm will be used to find the best answer.
- Technology: Learning to rank

Answer extraction

- From a given piece of text, extract the word or phrase which can answer the question exactly.
- The task could be implemented by finding the start word and the end word of the answer in the text.
- Answer extraction is used in Open Domain QA and Machine Reading Comprehension.

Answer extraction



Seo, Min Joon et al. Bidirectional Attention Flow for Machine Comprehension. CoRR abs/1611.01603, 2016

Database querying

- For NLIDB:
  - The user's question is converted to a database query (for example, a SQL statement) by the question understanding module.
  - The query is then sent to the database which returns a result.
  - The result is sent back to the user as the answer, possibly after post-processing by an anwer generation module.

Knowledgebase querying

- For KBQA or KGQA:
  - The user's question is converted to a knowledge base (or knowledge graph) query (for example, a SparQL statement) by the question understanding module.
  - The query is then sent to the knowledge base (or knowledge graph) which returns a result.
  - The result is sent back to the user as the answer, possibly after post-processing by an anwer generation module.

SparQL Query

SPARQL Query

Query:

```
PREFIX ocrer:   <http://purl.org/net/OCRe/research.owl#>
PREFIX ocresd:  <http://purl.org/net/OCRe/study_design.owl#>
PREFIX skos:    <http://www.w3.org/2004/02/skos/core#>
PREFIX vcard:   <http://www.w3.org/2006/vcard/ns#>
PREFIX vitro-public: <http://vitro.mannlib.cornell.edu/ns/vitro/public#>
PREFIX vivo:    <http://vivoweb.org/ontology/core#>
PREFIX scires:  <http://vivoweb.org/ontology/scientific-research#>
PREFIX core: <http://vivoweb.org/ontology/core#>

Select ?s ?p ?o
where {
  ?s a vivo:Relationship .
}
ORDER BY ?s
```

Format for SELECT and ASK query results:

◉ RS_TEXT ○ CSV ○ TSV ○ RS_XML ○ RS_JSON

Format for CONSTRUCT and DESCRIBE query results:

○ N-Triples ◉ RDF/XML ○ N3 ○ Turtle ○ JSON-LD

Run Query

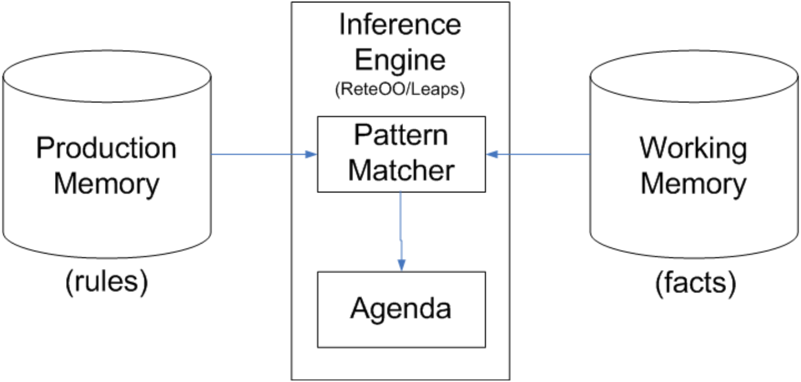
hrefhttps://wiki.lyrasis.org/display/VIVODOC110x/SPARQL+QueriesVIVO 1.10.x Documentation: SPARQL Queries

Rule-based inference engine

- Inference engine is a component of the system that applies logical rules to the knowledge base to deduce new information.
- The first inference engines were components of expert systems. The typical expert system consisted of a knowledge base and an inference engine.
- The knowledge base stored facts about the world.
- The inference engine applies logical rules to the knowledge base and deduced new knowledge.
- This process would iterate as each new fact in the knowledge base could trigger additional rules in the inference engine.

Wikipedia: Inference Engine

Rule-based inference engine



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

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))
...

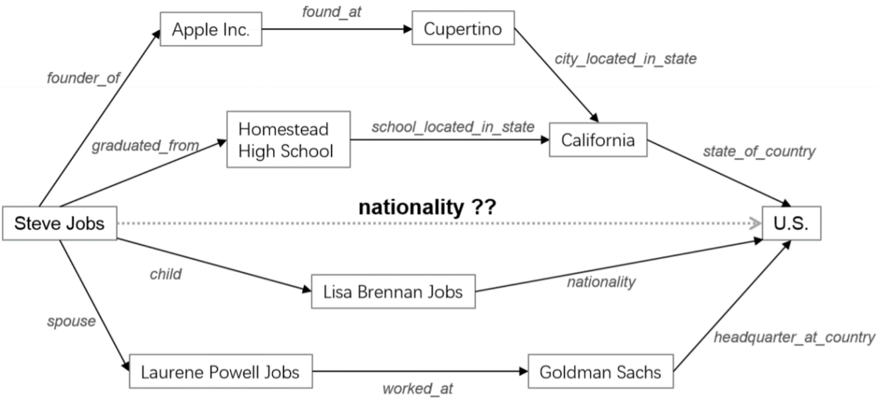
Fernandes et al., A rule-based system proposal to aid in the evaluation..., arXiv:1811.12454

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

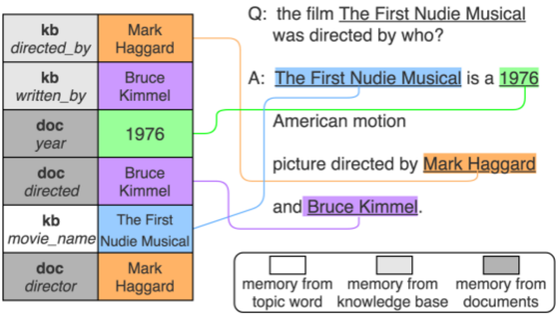
Inference over knowledge graphs



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.



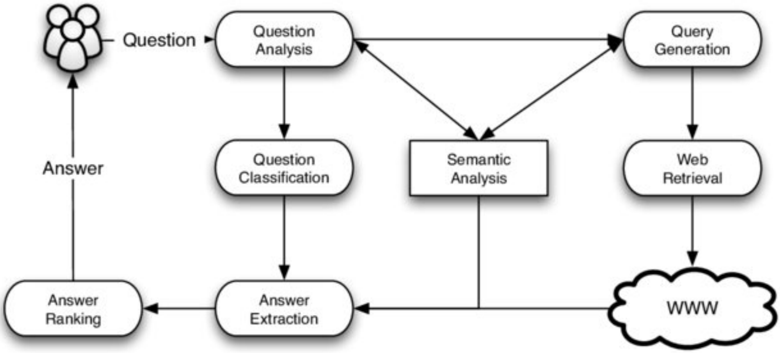
Yao Fu and Yansong Feng, Natural Answer Generation with Heterogeneous Memory, NAACL-HLT 2018

Content

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

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?

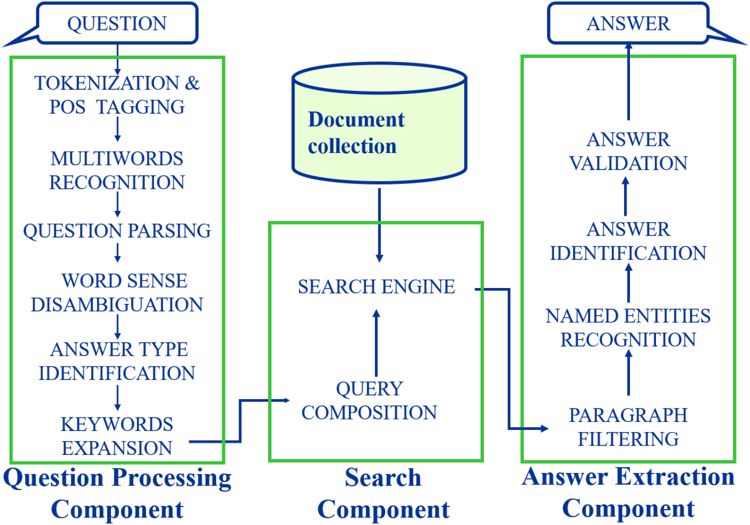
DIOGENE: Enola
PARAGRAPH: NYT19991001.0143

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)



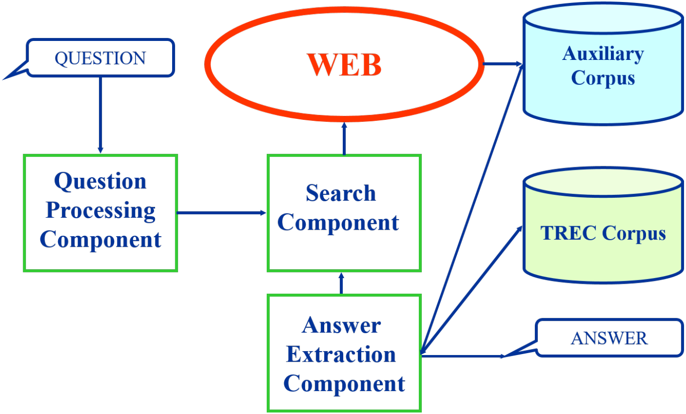
Bernardo Magnini, Open Domain Question Answering (slides), RANLP 2005

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)

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

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

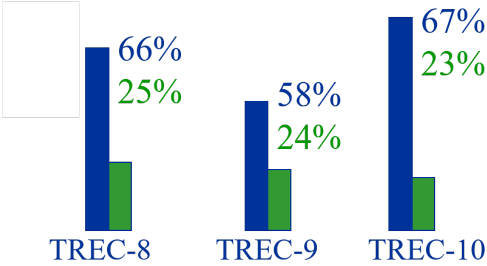
$$\frac{0 + ((0+0)/2) + (0+0+1)/3) + ((0+0+1+1)/4) + ((0+0+1+1+1)/5)}{5}$$

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!)

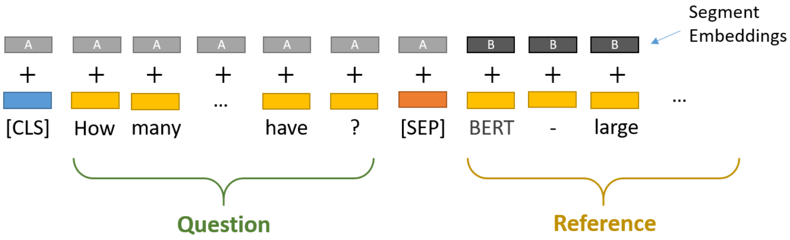
Thomas Lukasiewicz, Advanced Machine Learning: Deep Learning for NLP: Lecture 11: Question Answering, 2019

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.

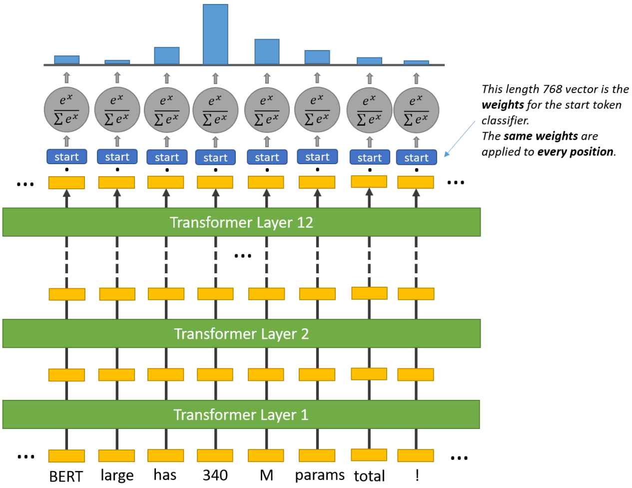
Chris McCormick, Question Answering with a Fine-Tuned BERT (Blog)

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)



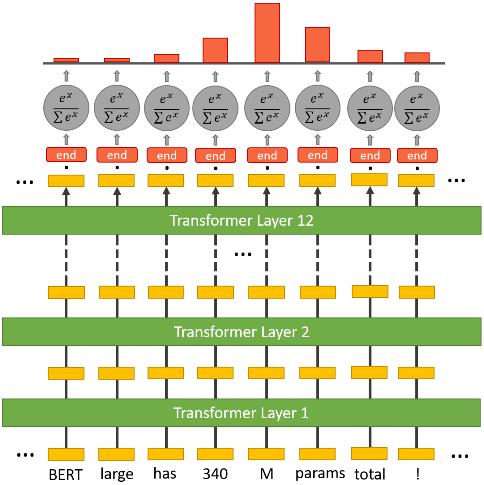
Chris McCormick, Question Answering with a Fine-Tuned BERT (Blog)

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

Content

Content

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:

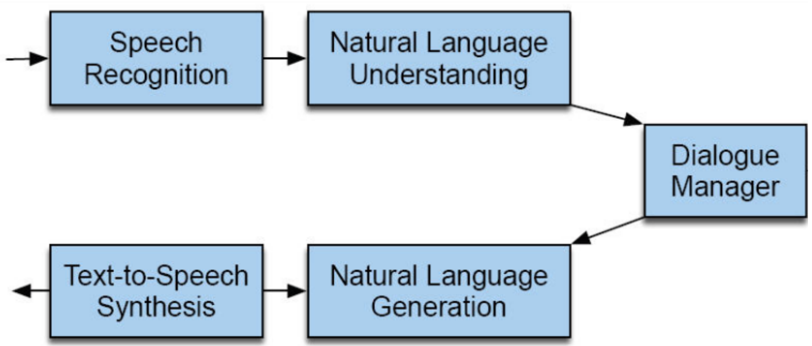
- Task-oriented (or goal-oriented) dialog systems
  - Help users to complete certain types of tasks.
  - Tasks (or domain knowledge) should be given in advanced, usually as a set of pre-defined intentions and slots.
  - Dialog sessions: the shorter is the better.
- Chitchat dialog systems (or social 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

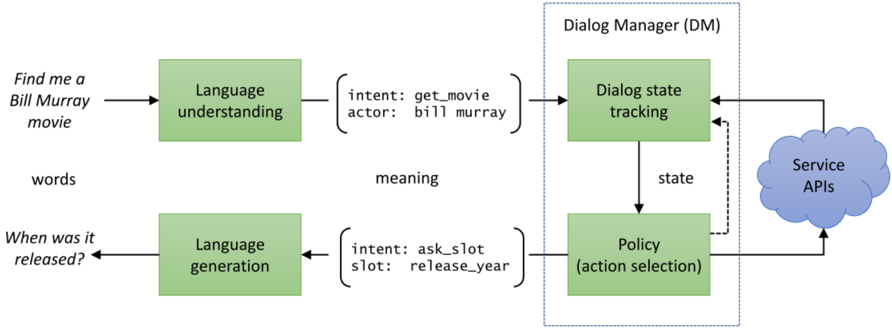
Dialog system structure



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**

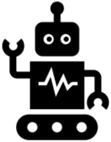
ConvLab: An Open Source Dialog System Platform

- An open source Multi-Domain End-to-End Dialog System Platform provided along with the 8th Dialog System Technology Challenge (DSTC8)

Fully annotate data
for training individual components or
end-to-end models with supervision

Speaker	Utterance	Annotation
User	am looking for a place to to stay that has cheap price range it should be in a type of hotel	**Dialog acts** {"Hotel-Inform": [["Pricerange", "cheap"]]}
System	Okay, do you have a specific area you want to stay in?	**State** {"hotel": {"name": "not mentioned", "area": "not mentioned", "parking": "not mentioned", "pricerange": "cheap", "stars": "not mentioned", "internet": "not mentioned", "type": "hotel"}} ... **Dialog acts** "Hotel-Request": [["Area", "?"]]
User	no, i just need to make sure it's cheap. oh, and i need parking	**Dialog acts** {"negate", "Hotel-Inform": [["Pricerange", "cheap"], ["Parking", "yes"]]}
System	I found 1 cheap hotel for you that includes parking. Do you like me to book it?	**State** {"hotel": {"name": "not mentioned", "area": "not mentioned", "parking": "yes", "pricerange": "cheap", "stars": "not mentioned", "internet": "not mentioned", "type": "hotel"}} ... **Dialog acts** {"Hotel-Inform": [["Price", "cheap"], ["Choice", "1"], ["Parking", "none"]]}

User Simulators
for reinforcement learning
1 rule-based simulator
2 data-driven simulators



SOTA Baselines
Multiple models for each component
Multiple end-to-end system recipes

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

ConvLab: An Open Source Dialog System Platform

- Paper: `https://arxiv.org/pdf/1904.08637.pdf`
- Codes and data: `https://github.com/ConvLab/ConvLab`
- A good start point to learn how to build an end-to-end dialog system.
- Here we use some examples from ConvLab to demonstrate the details of a task-oriented dialog system.

ConvLab: Dialog Acts

general-bye	Booking-Book	Hotel-Inform
general-greet	Booking-Inform	Hotel-NoOffer
general-reqmore	Booking-NoBook	Hotel-Recommend
general-thank	Booking-Request	Hotel-Request
general-welcome	Police-Inform	Hotel-Select
Attraction-Inform	Police-Request	Restaurant-Inform
Attraction-NoOffer	Train-Inform	Restaurant-NoOffer
Attraction-Recommend	Train-NoOffer	Restaurant-Recommend
Attraction-Request	Train-OfferBook	Restaurant-Request
Attraction-Select	Train-OfferBooked	Restaurant-Select
Hospital-Inform	Train-Request	Taxi-Inform
Hospital-Request	Train-Select	Taxi-Request

ConvLab: Dialog Slots

Fee	Ref	Day	none
Addr	Food	Name	Depart
Area	Type	Car	People
Stars	Price	Leave	Dest
Internet	Stay	Time	Parking
Department	Phone	Arrive	Open
Choice	Post	Ticket	Id

ConvLab: Training data (Multiwoz)

sessID	MsgID	Text	DialogAct
PMUL1032	0	What kind of attractions are available in the centre?	Attraction-Inform(Area=centre)
PMUL1032	1	There is the Holy Trinity Church on Market Street. It is free to get in.	Attraction-Inform(Addr=Market Street;Fee=free;Name=Holy Trinity Church)
PMUL1032	3	Sorry there are no listings for multiple sports,can I check in another area?	Attraction-NoOffer(Type=multiple sports)
PMUL1032	4	How about any that are about architecture?	Attraction-Inform(Type=architecture)
PMUL1032	8	Yes, may I have the phone number and postcode? Also, is there an entrance fee?	Attraction-Request(Post;Phone;Fee)
PMUL1032	9	The phone number is 01223332320 and the entrance fee is free. The post code is cb21tt.	Attraction-Inform(Fee=free;Phone=01223332320;Post=cb21tt)
PMUL1032	10	Thank you that is all I needed.	general-thank()
PMUL1033	0	I need to book a train to cambridge on Monday.	Train-Inform(Dest=cambridge)
PMUL1033	1	Where will you be departing from?	Train-Request(Depart)
PMUL1033	2	I will be departing out of Stevenage.	Train-Inform(Depart=stevenage)
PMUL1033	4	I want to arrive by 16:45.	Train-Inform(Arrive=16:45)
PMUL1033	6	Sorry, I looked at the calendar. I need a Thursday train, not a Monday. Can you please find a train on that day instead?	Train-Inform(Day=thursday)
PMUL1033	7	Okay, no problem. The TR1163 train leaves at 05:54. Will that work for you?	Train-Inform(Id=TR1163;Leave=05:54)
PMUL1033	8	What time will the train arrive in Cambridge?	Train-Inform(Dest=cambridge)
PMUL1033	9	It arrives at 06:43.	Train-Inform(Arrive=06:43)
PMUL1033	10	Ok please book that for 5 people.	Train-Inform(People=5)

Edit

a small piece

ConvLab: Distribution of Dialog Acts & Slots

User Dialog Act:		
Attraction-Inform		
	Area	2084
	Name	1447
	Type	2477
	none	269
Attraction-Request		
	Addr	1236
	Area	379
	Fee	1109
	Phone	1314
	Post	1401
	Type	258
Hospital-Inform		
	Department	95
	none	231

System Dialog Act:		
Attraction-Inform		
	Addr	2664
	Area	2365
	Choice	2241
	Fee	1922
	Name	3072
	Open	13
	Phone	1741
	Post	1579
	Price	48
	Type	2589
	none	28
Attraction-NoOffer		
	Addr	1
	Area	324

a small piece

NLU: Natural Language Understanding

- Tasks:
  - Intent (dialog act) detection: text classification
  - Slot filling: sequence labeling
- Technologies:
  - Intent detection: SVM, CNN, CNN-LSTM, BERT
  - Slot filling: CRF, Bi-LSTM, Bi-LSTM-CRF, ELMo/BERT
- Challenges:
  - Low resource
  - Domain adaptation
  - Out-of-domain problem

DST: Dialog State Tracking

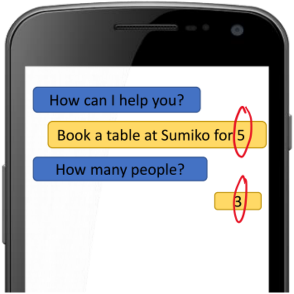
- Tasks:
  - Update the dialog states with the lastest input
- Technologies:
  - Rule-based method: when you have a high quality NLU module, a simple rule-based method to incorporate the NLU result with exisitng states (adding new slots or updating existing slots) will work well.
  - Neural method: it is a common practice to construct a neural DST module which has NLU included.

Belief state: distribution of dialog states

- Maintain a probabilistic distribution instead of a 1-best prediction for better robustness to LU errors or ambiguous input

Slot	Value
# people	5 (0.5)
time	5 (0.5)

Slot	Value
# people	3 (0.8)
time	5 (0.8)



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

Young, S., Gašić, M., Keizer, S., Mairesse, F., Schatzmann, J., Thomson, B. and Yu, K., 2010. The hidden information state model: A practical framework for POMDP-based spoken dialogue management. Computer Speech & Language, 24(2), pp.150-174.

Neural Belief Tracker: Data-Driven Dialogue State Tracking

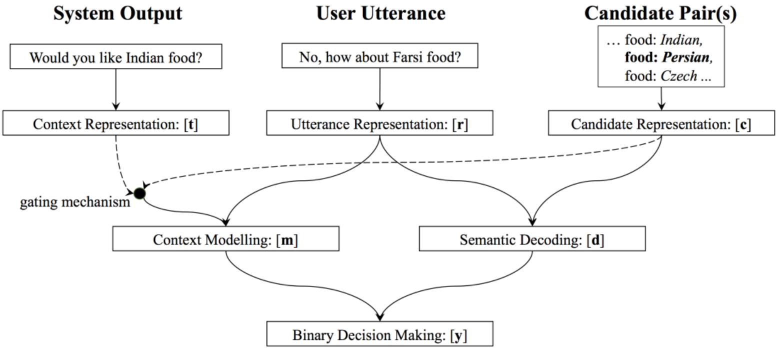


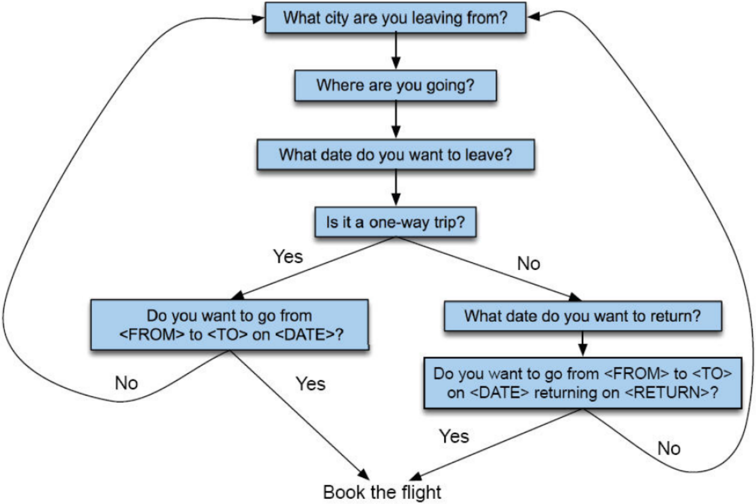
Figure 3: Architecture of the NBT Model. The implementation of the three representation learning subcomponents can be modified, as long as these produce adequate vector representations which the downstream model components can use to decide whether the current candidate slot-value pair was expressed in the user utterance (taking into account the preceding system act).

Mrkšić, N., Séaghdha, D.O., Wen, T.H., Thomson, B. and Young, S., 2016. Neural belief tracker: Data-driven dialogue state tracking. arXiv preprint arXiv:1606.03777.

DP: Dialog Policy

- Dialog policy decides what and how a taskoriented dialog system will respond, according to the current dialog states.
- Dialoy policy plays a vital role in delivering effective conversations.
- Approaches for dialog policy:
  - Rule-based policy
  - Supervised learning policy
  - Reinforcement learning policy

Rule-based policy



Original slides by Dan Jurafsky

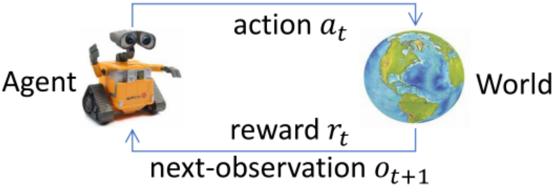
Rule-based policy

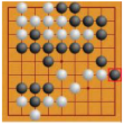
- Simple and straightforward. Easy to implement.
- System takes the control of the whole conversation.
  - However, in a real conversatipon, the user may not strictly follow the pre-defined flow. For example:
    - The user may prefer to give the details in a different order than the pre-defined one;
    - The user may give multiple details in one utterance;
    - The user may ask, confirm, or clarify something.
  - In such a scenario, even if the user's response is pertinent to the task, the system will not be able to complete the task.
- The conversation will be inflexible and unfriendly to users.

Policy by supervised learning

- The policy may be learned with a supervised learning methods.
- Given the current dialog states, learn a policy to predict the action (dialog act and slots) from the training data.
- The supervised learning for dialog policy is limited by the small size of the training data and the lack of exploration of the dialog state space.
- Since dialog systems interact with the enviroment constantly, reforcement learning would be a better choice in such a scenario.

Reinforcement Learning (RL)





Goal of RL

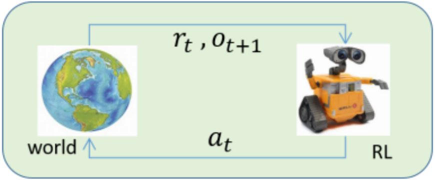
At each step t, given history so far s_t, take action a_t to maximize long-term reward ("return"):

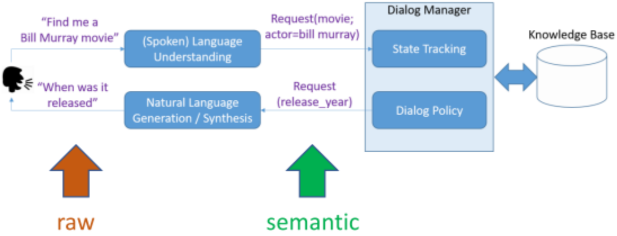
$$R_t = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \cdots$$

"Reinforcement Learning: An Introduction", 2nd ed., Sutton & Barto

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

Conversation as RL





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

Conversation as RL

- State and action
  - Raw representation
    (utterances in natural language form)
  - Semantic representation
    (intent-slot-value form)
- Reward
  - +10 upon successful termination
  - -10 upon unsuccessful termination
  - -1 per turn
  - ...

Early work:

Esther Levin, Roberto Pieraccini, and Wieland Eckert. A stochastic model of human-machine interaction for learning dialog strategies. IEEE Transactions on speech and audio processing 8.1 (2000): 11-23.

Pietquin, Olivier. A framework for unsupervised learning of dialogue strategies. Presses univ. de Louvain, 2005.

Jason D. Williams and Steve Young. Partially observable Markov decision processes for spoken dialog systems. Computer Speech & Language 21.2 (2007): 393-422.

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

Conversation as RL

- RL for dialog policy is very promising and has sparkled a lot of research:
  - Value-based RL
  - Policy-based RL
  - Model-based RL
- Despite of the success of RL in dialog system research, rule-based policy is still commonly used in industries because of its simplicity and being easy to handle.

User simulation

- One of the research lines is to use user simulators:
  - Train a user simulator off-line by rule-based methods or supervised learning on the training corpus;
  - Train the dialog policy on-line with the responses from the user simulator by reinforcement learning.
- ConvLab provides two types of user simulator:
  - rule-based simulators, and
  - data-driven simulators.

NLG: Natural Language Generation

- NLG module generate the system utterance according to the speech act (or action, including the intent and slot-value pairs), which is provided by the dialoy policy module.
- For task-oriented dialog systems, if the task is not too complex, a rule-based generation should work well.
- In recent years, deep learning based NLG for dialog systems has become popular and can generate texts in rather good quality.

End-to-End training for dialog systems

- The training of a pipeline dialog system is complex and hard to adopt to new domain.
- End-to-end training for dialog systems has been proposed and become a popular paradigm.
- The data for end-to-end training contains not only the dialog utterances from the user and the dialog system, but also the dialog acts and slot-value pairs for each utterance, as we have seen in the ConvLab system.

End-to-End training for dialog systems

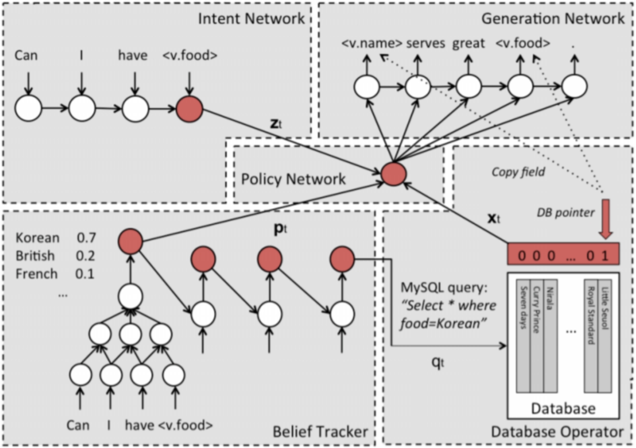


Figure 1: The proposed end-to-end trainable dialogue system framework

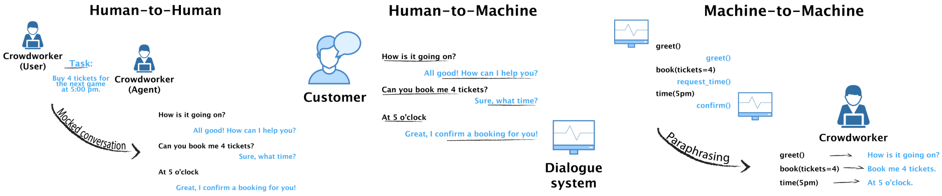
Tsung-Hsien Wen, et al. A Network-based End-to-End Trainable Task-oriented Dialogue System. EACL 2017.

Data Collection and Wizard-of-Oz

- To collect dialog data from real human-to-human conversation is really difficult.
- A Wizard-of-Oz (WoZ) paradigm is proposed to solve this problem:
  - a human (e.g., a crowdworker) takes the role of a user with a specic task in mind: for instance, she wants to book a restaurant through a call center, or has a question for a customer service.
  - The role of the system agent (so called the wizard) is then played by another crowdworker who has access to the knowledge (e.g., databases, FAQs, the user's history) required to complete the task.
  - An actual conversation is set within a particular domain and “mocked” between the two parties.

Data Collection and Wizard-of-Oz

- WoZ has been used to create two largest research datasets for task-oriented dialogue today: Stanford multi-domain dataset (SMD) and Multi-Domain Wizard-of-Oz (MultiWOZ)
- Improved versions of WoZ experiments have been proposed and used to collect dialog data:



The Magic Triangle of Dialogue Data Collection, PolyAI

Evaluation for task-oriented dialogs

- Many metrics can be applied for evaluation of task-oriented dialog system, which are mainly categorized as:
  - Human evaluation:
    - Evaluation done by real human users.
    - Reliable but expensive.
    - Metrics: Task success rate, dialog length, irrelevant turn rate, redundant turn rate, user satisfaction score, etc.
  - Automatic evaluation:
    - Evaluation by the user simulator.
    - Cheap but not so reliable.
    - Metrics: Task success rate, dialog length, average rewards, etc.

Content

Chitchat dialog systems

- A chitchat dialog system (social bot) is expected to chat with the user freely on any topics.
- It had been an extremely difficult task and not practically feasible until the neural sequence-to-sequence was applied:
  - L. Shang, Z. Lu, and H. Li. Neural responding machine for short-text conversation. ACL-IJCNLP 2015.
  - O. Vinyals and Q. Le. A neural conversational model. arXiv preprint arXiv:1506.05869, 2015.

Seq2seq generation for chitchat dialogs

- The sequence-to-sequence model used for chitchat dialogs is similar to neural machine translation models:
  - An encoder takes the dialog history as the input and encodes it into an internal representation;
  - An decocer takes the internal representation as the input and decodes it into an output sequence as the response.
- Huge open domain conversation data can be collected from social media, which provide almost unlimited resource to build open domain chitchat dialog systems.

Seq2seq generation for chitchat dialogs

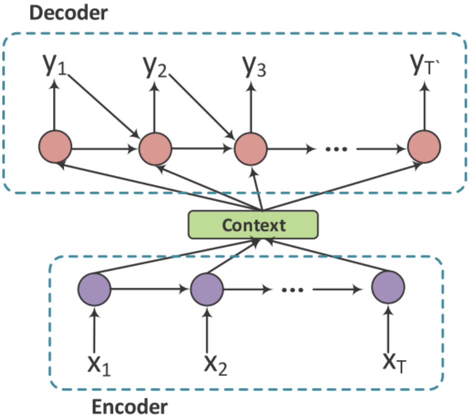


Figure 2: An Illustration of the Encoder-Decoder Model.

Hongshen Chen et al., A Survey on Dialogue Systems: Recent Advances and New Frontiers, ACM SIGKDD Explorations Newsletter 19.2 (2017)

A piece of dialog generated by a seq2seq model

Neural model learns the general shape of conversation, and the system output is situationally appropriate and coherent…

H:	Hi
M:	**How are you?**
H:	Not bad
M:	**What's wrong?**
H:	Nothing really
M:	**Wanna talk about it? I'm here if you wanna talk x**
H:	Well, I'm just tired
M:	**Me too, I'm here if you wanna talk**
H:	Thanks, what are you doing?
M:	**You're welcome and I'm watching TV and you?**
H:	On Netflix?
M:	**Yeah, I'm watching it on Netflix.**
H:	What are you watching?
M:	**A TV show.**
H:	Nice, did you have dinner already?
M:	**No, not yet. what about you?**

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

Problems for seq2seq dialog generation

- Although neural seq2seq has made great progress in chitchat dialog generation, there are still many problems for this method:
  - Universal response: the system frequently generate universal response like "I don't know.", "Ok", etc.
  - Inconsistant: the system often provides contradictory information in the same session: "I'm 14.", "I'm 16."
  - Long context: it is difficult to encode very long dialog history.
  - External knowledge: it is difficult to incorporate external knowledge into the dialog generation process.
  - Persona: it is hard to assign a specific personality to the system.
- Intensive research has been conducted in this research area in recent years.

Retrieval-based chitchat generation

- Another effective to produce natural chitchat dialog is to use retrieval-based method, if the dialog data is large enough.
- Retrieval-based method and generation-based method are complementary to each other and are often used togethor to form a hybrid method.

Evaluation for chitchat dialogs

- Automatic evaluation of chitchat dialog remains an open question.
- Evaluation metrics borrowed from machine translation and text summarizaiton like BLUE, METEOR and ROUGH are offen used, although not satisfactory.
- New metrics using word embeddings or neural networks are also proposed.

Content

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