Conversational Modeling

Ramaseshan Ramachandran

TYPES OF CONVERSATIONS

- Threaded
  - Twitter, Facebook, email
- Short-text conversation
  - Google help desk, Microsoft Virtual Agent, etc. where the interactions are $\geq 1 < 3$
- Task-oriented conversation
  - Siri, Cortana, Google Home, Alexa, help desk, etc.- to get information from the user to help complete the task
- Chit-chat or open conversation - unstructured conversations on any topic
- Question answering

Year	Event
1950	Turing Test
1955	AI Born
1964	ELIZA
2011	Siri
2011	IBM Watson
2014	Alexa
2016	Google Home

Modeling conversation is one of the active research problems in AI

Natural language conversation involves language understanding, reasoning, and the utilization of common sense knowledge

The goal is to build a conversational model that generates the responses automatically and these responses are linguistically indistinguishable from human responses thereby passing the *Turing Test*

A true test for machine intelligence

Can a machine be linguistically indistinguishable from a human?

icons from www.flaticon.com
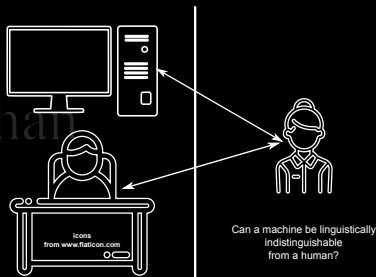
Would you like some coffee?

* Yes, please

Mega, would you like to dance?

* Is the floor slippery?

* No, it's fine

* **Teacher**: Will you tell us the answer to question four?

* **Mike**: Is that one on page (...) six or seven? Then I'd be happy to

* **Teacher**: Six

* **Mike**: Oh, okay. The answer is factorial two

CONVERSATION ANALYSIS

▶ Understanding what is **NOT** said

▶ Analysis of the language beyond sentence

▶ Identification of the relationship among all of the contexts across sentence boundaries

▶ Consists of two parts - **Representation and Conditions**

   1. Representation - a set of referents representing the entities which are under discussion
   2. Conditions - a set of conditions representing the entities
      Example
      A farmer owns a donkey
      $[x, y : farmer(x), donkey(y), owns(x, y)]$

▶ Relationship - how two segments of discourse are logically connected to each other

GOALS OF THE CONVERSATIONAL AGENT

► Machine Comprehension of text
► Learn conversations
► Learn to summarize or encode conversation
► Predict the meaning of the conversation
► Produce substantiative and useful conversation
► Produce a sentence with a good fluency
► Detect hypocrisy, abuses, flirting, threats, etc.
► Learn task-oriented conversation
► Learn typical patterns of dialog discourse
► Learn to identify questions
► Learn Question-Answering

ELIZA - DR. WEIZENBAUM

The first AI based conversation engine

- ▶ Idea is to simulate methods used by a psychologist that involve drawing the patient out by reflecting patient's statements back at them
- ▶ Human-like, but does not understand the query/conversation
- ▶ Uses pattern matching with some rules - insertion, substitution and deletion
- ▶ No built in context understanding mechanism

ELIZA-A TYPICAL CONVERSATION

Men are all alike.

IN WHAT WAY

They're always bugging us about something or other.

CAN YOU THINK OF A SPECIFIC EXAMPLE [1]

Well, my boyfriend made me come here.

YOUR BOYFRIEND MAKE YOU COME HERE

he says i'm depressed much of the time.

I AM SORRY TO HEAR YOU ARE DEPRESSED

it's me. I am not happy.

DO YOU THINK COMING HERE WILL, HELP YOU NOT TO BE HAPPY

I need some help, that much seems certain

WHAT WOULD IT MEAN TO YOU IF YOU GOT SOME HELP [2]

---

[1] The capitalized conversations in yellow are the machine responses

[2] Source:J.Weizenbaum, "ELIZA A Computer Program For the Study of Natural Language Communication Between Man And Machine",vol9, Jan. 1966

OK, let's start over. Please describe your issue and I'll look for the best solution.

How to find office licence key?

Which one did you mean?

How to find your Office product key

How to activate or find product key for Office Home Use Program (HUP)

None of the above

How to find your Office product key

Choose your Office version:

Office 365

Office 2019, Office 2016, or Office 2013

Office 2011 for Mac

Office 2010

Office 2007

Office 2011 for Mac

Do you need help finding your key or did you lose the key?
Find product key
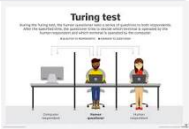Lost product key

Find product key

An Office for Mac 2011 product key is a 25-character code used to activate Office for Mac 2011. Where to find your product key depends on how you got your copy of Office for Mac 2011:

- **From an authorized retailer**. The product key is on a label, card, or sticker inside the box or case that Office came in. If you downloaded Office from an online store, the product key should be in your email receipt.
  - **IMPORTANT**: If Office came on a product key card with a 27-character alpha-numeric PIN, you'll need to contact support to redeem the PIN.
- **A digital copy from a Microsoft website**. The product key is in the confirmation email you received after buying it. To find the email, try the following:
  - Check your spam, bulk, or junk mail folders for the email.
  - Try searching your email for the word *Microsoft*.
  - If you have multiple email accounts, check the email account that you provided when you purchased Office for Mac 2011 online.

Source Credit: https://support.microsoft.com/en-in/contact/virtual-agent/
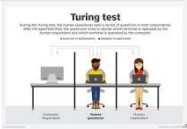
QUESTION ANSWERING - EXAMPLE

Highlevel architecture

Sample Jeopardy! game board
Image Source: https://commons.wikimedia.org/
wiki/File:Jeopardy!_game_board.png

https://www.aaai.org/Magazine/Watson/watson.php
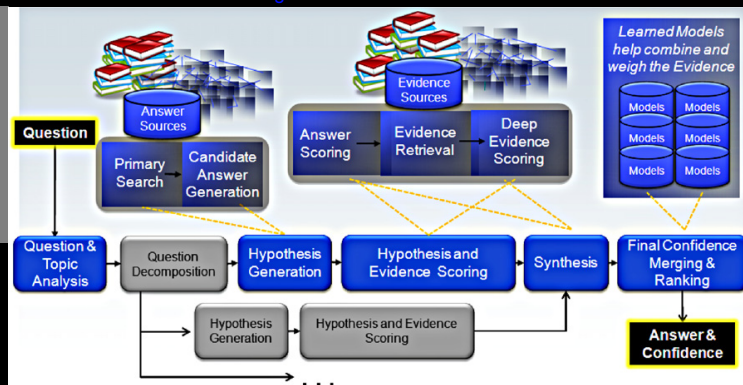
- Retrieval-based Approach
  - Pick a suitable response based on how many times a particular response was selected for similar questions
  - Using matching features of question and the response
    - The use of matching features alone will not suffice
- Statistical Machine Translation approach
  - This approach treats this as a translation problem in which the model is trained on the parallel corpus of question and response pairs

- IR based mostly used in the short-text conversation[3]

- The corpus contains different pairs of post-comments or question answers

* Given a question, and the set of documents, the task is to find the answer from the span of text from extracted paragraphs

For every given query q, there could be zero or more post-comment pairs (p, r) The best response to the query q is picked up based on the ranks of the retrieved

pairs using

$$\hat{r} = \underset{(p,r)}{\arg\max} \, Score(q, (p, r)) \qquad (1)$$

where $Score(.)$ is the sum of all score of the features

$$Score(q, (p, r)) = \sum_{i \in \Omega} w_i \phi_i(q, r) \qquad (2)$$

where $\phi_i(.)$ and w_i are the score and weight of the ith feature and Ω is the total number of features, respectively. Here the features could be TF*IDF of the word found in the {q,(p,r)}

[3]Zongcheng Jia, Zhengdong Lub, Hang Li, An Information Retrieval Approach to Short Text Conversation, arXiv:1408.6988v1 [cs.IR] 29 Aug 2014

IR BASED MODELING - ARCHITECTURE

▶ Query-Post Similarity: Here the similarity between the query and the candidate responses are computed using similarity measures such as cosine similarity

$$Similarity(q, p) = \frac{q^T p}{||q|| \cdot ||p||} \quad (4)$$

These similarity measures are proposed with the assumption that the there is some alignment of variables between query and posts and query and responses
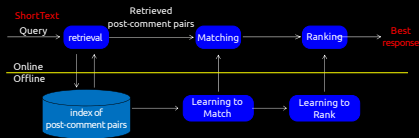
▶ Query-Response Similarity: Here the similarity between the query and the candidate responses are computed using similarity measures such as cosine similarity

$$Similarity(q, r) = \frac{q^T r}{||q|| \cdot ||r||} \quad (3)$$

LEARNING TO RANK

IR BASED CONVERSATIONS
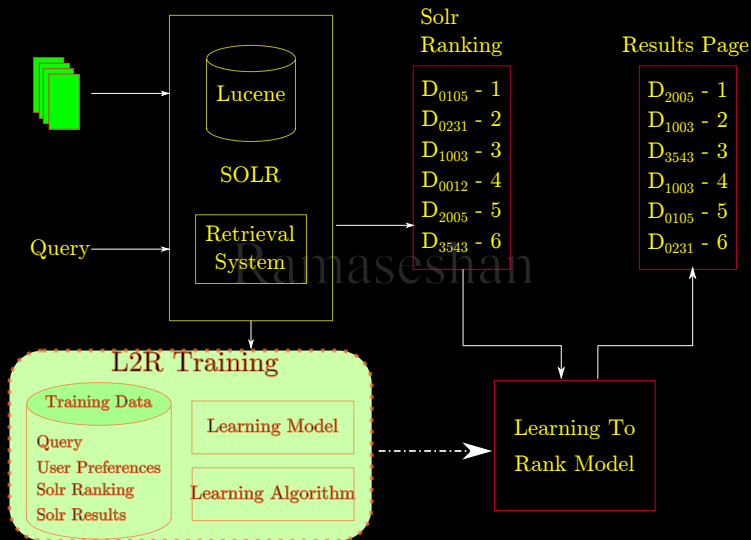
The main drawbacks of the retrieval-based method are the following

► The Post,responses pairs are canned and it is hard to customize for the particular text or requirement from the task, e.g., style or attitude

► The use of matching features alone is usually not sufficient for distinguishing positive responses from negative ones, even after time consuming feature engineering. (e.g., a penalty due to mismatched named entities is difficult to be incorporated into the model)

QUESTION ANSWERING - INTRODUCTION

An application of short Post-Response is Question Answering system, such as IBM Watson (Jeopardy)

In this case most of the candidate responses are answers for factoid questions

- ▶ Open domain question answering has become important research area in natural language processing
- ▶ Tougher than common search engine tasks
    - ▶ Finding accurate and concise answers to questions rather than a set of relevant document
- ▶ Simple term-based retrieval won't be enough
- ▶ **Type** of the sought after answer should be known to retrieve accurate answers

QUESTION ANSWERING - SAMPLES

Question	Hierarchy	Type
What is RNN?	Abbreviation	Expansion
Where is the big temple in India located?	Location	City
Who was the president of India in 2006?	Human	Person
Name the currency used in China	Entity	Currency
How far away is the moon?	Numeric	Distance
What is the chemical symbol for oxygen?	Entity	Symbol
What is a prism?	Description	Definition
Why is the sun yellow?	Description	Reason
When did CV Raman receive his Nobel Prize?	Numeric	Year

Most questions could be classified in to 6 major classes[4] - ABBREVIATION, ENTITY,DESCRIPTION, HUMAN, LOCATION and NUMERIC VALUE and around 50 fine-grained types.

---

[4]Xin Li, Dan Roth, Learning Question Classifiers

FEATURE SPACE

▶ Words

▶ Part of Speech (POS) tags

▶ Chunks(non-overlapping phrases)

▶ Named entities

▶ Head chunks(using POS - first noun chunk in a sentence)[5]

    A/DET trip/NOUN to/ADP Cape/NOUN Carnival/NOUN ,/PUNCT FL/NOUN ,/PUNCT

    takes/VERB 10/NUM hours/NOUN ./PUNCT The/DET distance/NOUN is/VERB 816/NUM km/NOUN

    ./PUNCT Calculate/VERB the/DET average/ADJ speed/NOUN

▶ Semantically related words (words that often occur with a specific question class - How far, How high, How long)

---

[5]A trip to Cape Carnival, FL, takes 10 hours. The distance is 816 km. Calculate the average speed

QUESTION TYPOLOGY RULES

Simple rules could be defined to classify questions
For example,

1. $if\ QuestionStartsWith(who)\ or\ QuestionStartsWith(whom)$
   $TopHierarchy \leftarrow HUMAN$
   $Class \leftarrow PERSON$
   fi

2. $if\ QuestionStartsWith(where)$
   $TopHierarchy \leftarrow LOCATION$
   $Class \leftarrow CITY$
   fi

   If a query contains Which or What, then the head noun phrase
   determines the class, as for What X questions
   What is a prism?

Given the list of classes and the features for each of the question, it is easy to calculate the probability distribution of classes for the given question

The probability density is

$$P = [p_1, p_2, \ldots, p_n] \qquad (5)$$

and the corresponding class labels are

$$C = [c_1, c_2, \ldots, c_n] \qquad (6)$$

p_is are obtained by employing Naive-Bayes algorithm
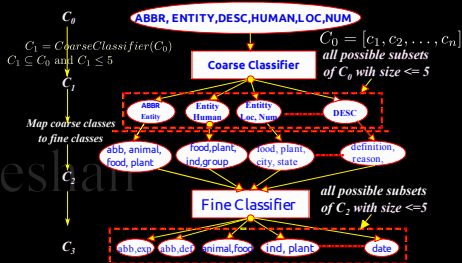
Figure source:Xin Li, Dan Roth, Learning Question Classifiers

ANSWER EXTRACTION

The important phase ⌣ in the QA system

Span Labeling: The span of text (tokens) that contains the answer. The task of finding the span of text is known as Span Labeling

Modern approaches combine a IR-based component based on bigram hashing and TF*IDF matching and a multi-layer recurrent neural network model trained to detect answers [6]

Emerging systems are designed as reading comprehension systems

Query

Question
Processing

Question
Type
Classification

Document and Passage Retrieval

Answer
Types
Answer
Merging
Answer
Score
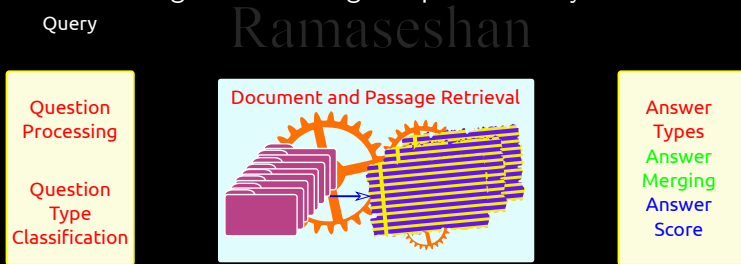
Answer

---
[6]Danqi Chen et al, Reading Wikipedia to Answer Open-Domain Questions

DOCUMENT RETRIEVER

▶ Using a typical Term-Document and the retrieval operations on the Term-Document matrix

▶ Using Inverted Indexing approach used in SOLR/Elastic search

▶ Using LSA

▶ Combination of the above with n-grams

▶ Using a ranking model to retrieve top 5-10 documents

▶ Use an answer encoder to find similar representations in the documents - Use of RNN

Who is CV Raman?

Sir CV Raman (7 November 1888-21 November 1970) was an Indian physicist born in the former Madras Province in India (presently the state of Tamil Nadu), who carried out ground-breaking work in the field of light scattering, which earned him the 1930 Nobel Prize for Physics. He discovered that when light traverses a transparent material, some of the deflected light changes wavelength and amplitude. This phenomenon, subsequently known as Raman scattering, results from the Raman effect[4] In 1954, the Indian government honored him with India's highest civilian award, the Bharat Ratna [5][6]

What is the invention of CV Raman?

Sir CV Raman (7 November 1888-21 November 1970) was an Indian physicist born in the former Madras Province in India (presently the state of Tamil Nadu), who carried out groundbreaking work in the field of light scattering, which earned him the 1930 Nobel Prize for Physics. He discovered that when light traverses a transparent material, some of the deflected light changes wavelength and amplitude. This phenomenon, subsequently known as Raman scattering, results from the Raman effect[4] In 1954, the Indian government honored him with India's highest civilian award, the Bharat Ratna [5][6]

FEATURES FOR ANSWERS

▶ Phrase matches keywords/patterns of question and the paragraph

▶ Count of terms that match question and potential paragraphs

▶ Cosine similarity

▶ Pattern matching using trained ANNs

▶ Probabilistic methods using alignment methods

Stanford Question Answering Dataset (SQuAD)

▶ Reading Comprehension Data set

▶ 87000 examples for training and 10000 examples for development

▶ All questions and answers are composed by humans through crowd sourcing

▶ The span of text is provided for all questions that could be answered

Steam_engine
The Stanford Question Answering Dataset

Steam engines are external combustion engines, where the working fluid is separate from the combustion products. Non-combustion heat sources such as solar power, nuclear power or geothermal energy may be used. The ideal thermodynamic cycle used to analyze this process is called the Rankine cycle. In the cycle, water is heated and transforms into steam within a boiler operating at a high pressure. When expanded through pistons or turbines, mechanical work is done. The reduced-pressure steam is then condensed and pumped back into the boiler.

Along with geothermal and nuclear, what is a notable non-combustion heat source?
Ground Truth Answers: solar solar power solar power, nuclear power or geothermal energy solar
Prediction: solar power

What ideal thermodynamic cycle analyzes the process by which steam engines work?
Ground Truth Answers: Rankine Rankine cycle Rankine cycle Rankine cycle
Prediction: Rankine cycle

In the Rankine cycle, what does water turn into when heated?
Ground Truth Answers: steam steam steam steam
Prediction: steam

At what pressure is water heated in the Rankine cycle?
Ground Truth Answers: high high high pressure high
Prediction: high

Datasets used: Stanford Question Answering Dataset-SQuAD, CuratedTREC, WebQuestions and WikiMovies

QUESTION ENCODING

▶ A question encoder creates weighted sum of all the words (q_i) in a question.

▶ The word embedding of each word in the question is fed to an RNN encoder

▶ For every time state, q_i, a hidden $\mathbf{q}_i$ is output from the hidden unit.

▶ For all the time states, a weighted sum $\mathbf{q}$ and a single embedding of the question is the output - $\mathbf{q} = [\mathbf{q}_1, \mathbf{q}_2, \mathbf{q}_3, \ldots \mathbf{q}_l]$

$$\mathbf{q} = \sum_j b_j q_j \tag{7}$$

$$b_j = \frac{\exp(\mathbf{w}.\mathbf{q}_j)}{\sum_i^t \exp(\mathbf{w}.\mathbf{q}_i)} \tag{8}$$

where $\mathbf{w}$ is the weight vector to be learned
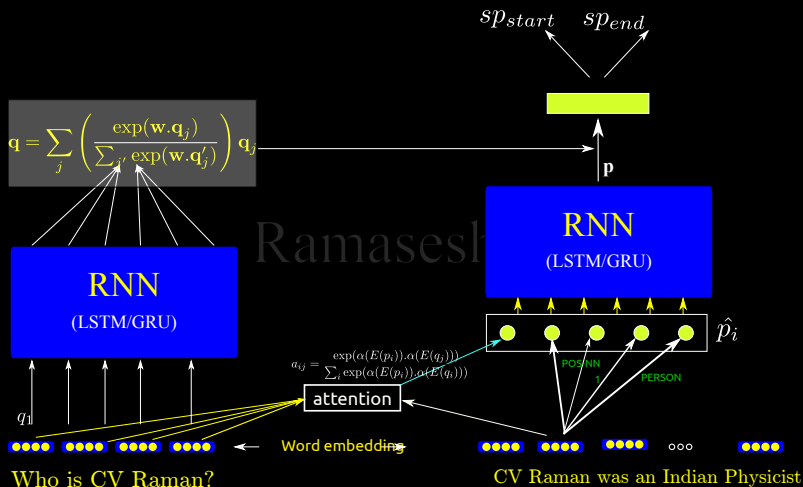
PARAGRAPH ENCODING

Let $q = (q_1, q_2, \ldots, q_n)$ be the question with n tokens

Let $\mathbf{p} = (\mathbf{p_1}, \mathbf{p_2}, \ldots, \mathbf{p_m})$ be the encoded paragraphs of $\hat{p} = (\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_m)$ and $\hat{p}_i$ represent the following:

1. The embedding of the word $f_1 = E(p_i)$

2. p_i can be matched exactly by one question word $f_2 = \mathbb{1}(p_i \in q_i)$

3. Token feature such as POS, NER, TF/TF*IDF - $f_{features}$

4. Aligned question embedding $f_{align}(p_i) = \sum_j a_{ij} E(q_j)$, where a_{ij} captures the similarity between p_i and q_j

$$a_{ij} = \frac{\exp(\alpha(E(p_i)).\alpha(E(q_j)))}{\sum_{j'} \exp(\alpha(E(p_i)).\alpha(E(q'_j)))} \tag{9}$$

$\alpha(.)$ is a single dense layer with ReLU nonlinearity. Compared to the exact match features, these features add soft alignments between similar but non-identical words(e.g.,car and vehicle)

$$sp_{start} \quad sp_{end}$$

$$\mathbf{q} = \sum_j \left(\frac{\exp(\mathbf{w}.\mathbf{q}_j)}{\sum_{j'} \exp(\mathbf{w}.\mathbf{q}'_j)} \right) \mathbf{q}_j$$

RNN
(LSTM/GRU)

RNN
(LSTM/GRU)

$\mathbf{p}$

$\hat{p}_i$

POS NN

PERSON

$$a_{ij} = \frac{\exp(\alpha(E(p_i)).\alpha(E(q_j)))}{\sum_i \exp(\alpha(E(p_i)).\alpha(E(q_i)))}$$

attention

q_1

Word embedding

Who is CV Raman?

CV Raman was an Indian Physicist

▶ The goal is to predict the span of tokens that is most likely the correct answer

▶ The RNN is trained using paragraph vectors $(\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_m)$ and question vector $\mathbf{q}$ to predict the span (sp_{start}, sp_{end})

▶ A bilinear attention layer $\mathbf{W}$ is used to predict instead of a simple similarity measure as follows:

$$sp_{start_i} \propto \exp(p_i \mathbf{W} \mathbf{q}) \tag{10}$$

$$sp_{end_i} \propto \exp(p_i \mathbf{W} \mathbf{q}) \tag{11}$$

▶ During prediction, the best span from $token_i$ to $token_{i'}$ such that $i \leq i' \leq i + 15$ and $sp_{start}(i) \times sp_{end}(i')$ is maximized.

EXPERIMENTS

▶ 3-layer bidirectional LSTMs with h = 128 hidden units for both paragraph and question encoding

▶ Stanford CoreNLP toolkit for tokenization and also generating lemma, part-of-speech, and named entity tags

Features	F1
Full	78.8
No f_{token}	78.0 (-0.8)
No f_{exact_match}	77.3 (-1.5)
No $f_{aligned}$	77.3 (-1.5)
No $f_{aligned}$ and $f_{e}xact_{m}atch$	59.4 (-19.4)

Most of the researchers use $F1$ score It is a weighted harmonic mean of *Precision* and *Recall* given by the relation:

$$F_\beta = \frac{(\beta^2+1)PR}{\beta^2 P + R}, where, \beta^2 = \frac{1-\alpha}{\alpha} \qquad (12)$$

where $\alpha \in \{0,1\}$ and $\beta \in \{0,\infty\}$. When $\alpha = \frac{1}{2}$ or $\beta = 1$, it is a balanced measure that gives equal weights to *Precision* and *Recall*

$$F_{\beta=1} = F_1 = \frac{2PR}{P+R} \qquad (13)$$

	Relevant	Not relevant
Retrieved	TP	FP
Not Retrieved	FN	TN

$$Precision = \frac{\#\,of\,relevant\,items}{\#\,of\,retrieved\,items} \qquad (14)$$

$$Recall = \frac{\#\,of\,relevant\,items\,retrieved}{\#,of\,Relevant\,items} \qquad (15)$$

$$Precision = TP/(TP+FP) \qquad (16)$$

$$Recall = TP/(TP+FN) \qquad (17)$$