Logistics

• Project progress updates 03/29
  • short presentations
UNIT 2

• Low-Resource NLP
• Summarization
• Dialog Systems
• Question Answering
Question Answering

Goal is to build systems that **automatically** answer questions posed by humans in a **natural language**
One of the oldest NLP tasks [Simmons et al. 1964]
Artists use color to create patterns. Color can also show different moods. Bright colors make us feel happy and energetic. Dark colors make us feel calm or sad.

The primary colors are red, yellow, and blue. They are the colors that can be mixed together to make different colors. Mixing two primary colors makes a secondary color. The secondary colors are orange, green, and violet (purple). Orange is made by mixing yellow and red. Green is made by mixing yellow and blue. Violet is made by mixing red and blue. Intermediate colors can be made by mixing a primary and a secondary color together. Some intermediate colors are blue violet and red orange. Black, white, and gray are special colors. They are called neutral colors.

Colors have been organized into a color wheel. It shows the three primary colors, the three secondary colors, and the six intermediate colors. Artists use the color wheel. It helps them know which colors they want to use together in their artwork.

1) What kinds of colors make us feel calm?
   - dark colors

2) What kinds of colors make us feel like we have lots of energy?
   - bright colors

3) What are the primary colors?
   - red, blue and yellow

4) What are the secondary colors?
   - green, orange and violet (purple)

5) What tool do artists use to organize all the colors?
   - a color wheel
Gagarin

April 12, 1961, Gagarin became the first person in space, making a 108-minute orbital flight in his Vostok 1 spacecraft. The 27-year-old cosmonaut traveled 327 kilometers above the Earth and orbited the planet at a speed of 27,400 kilometers per hour.

Apr 13, 2020

spacecenter.org > mission-monday-five-fast-facts-about-th...
• Testing language understanding
  • Understand the question
  • Relate to a position in document
  • Extract the answer
Lehnert 1977: “Since questions can be devised to query any aspect of text comprehension, the ability to answer questions is the strongest possible demonstration of understanding.”

• Addressing human information needs
  • Talking to a virtual assistant
  • Interacting with a search engine
  • Querying a database
NLP task $\Leftrightarrow$ QA task

- NLP tasks can be reduced to machine reading
- Given passage extract information [Levy et al. 2017]

(Barack Obama, educated_at, ?)

Question: Where did Barack Obama graduate from?

Passage: Obama was born in Honolulu, Hawaii. After graduating from Columbia University in 1983, he worked as a community organizer in Chicago.
Human Need for Information

Smart Speaker Use Case Frequency January 2020

- **Listen to streaming music service**: 88.7% (EVER TRIED), 73.6% (MONTHLY), 39.8% (DAILY)
- **Ask a question**: 83.1% (EVER TRIED), 66.2% (MONTHLY), 29.4% (DAILY)
- **Check the weather**: 77.1% (EVER TRIED), 59.8% (MONTHLY), 33.9% (DAILY)
- **Set a timer**: 64.5% (EVER TRIED), 52.4% (MONTHLY), 20.3% (DAILY)
- **Set an alarm**: 59.8% (EVER TRIED), 45.6% (MONTHLY), 26.3% (DAILY)
- **Listen to the radio**: 59.8% (EVER TRIED), 42.6% (MONTHLY), 19.0% (DAILY)
- **Listen to News / Sports**: 50.6% (EVER TRIED), 37.7% (MONTHLY), 16.9% (DAILY)
- **Use a favorite Alexa skill or Google Action**: 47.9% (EVER TRIED), 34.0% (MONTHLY), 16.4% (DAILY)
- **Play game or answer trivia**: 46.1% (EVER TRIED), 27.7% (MONTHLY), 9.0% (DAILY)

**Other Use Cases**

- **Listen to Podcast or other talk formats**: 44.9% (EVER TRIED), 32.0% (MONTHLY), 11.4% (DAILY)
- **Control smart home devices**: 43.4% (EVER TRIED), 31.9% (MONTHLY), 24.5% (DAILY)
- **Find a recipe or cooking instructions**: 42.3% (EVER TRIED), 26.0% (MONTHLY), 5.4% (DAILY)
- **Call someone**: 40.2% (EVER TRIED), 21.2% (MONTHLY), 9.5% (DAILY)
- **Search for product information**: 38.2% (EVER TRIED), 27.9% (MONTHLY), 7.3% (DAILY)
- **Check traffic / directions**: 35.1% (EVER TRIED), 23.7% (MONTHLY), 11.1% (DAILY)
- **Access my calendar**: 32.1% (EVER TRIED), 19.0% (MONTHLY), 9.5% (DAILY)
- **Send a text message**: 27.8% (EVER TRIED), 14.9% (MONTHLY), 6.7% (DAILY)
- **Make a purchase**: 25.2% (EVER TRIED), 4.9% (MONTHLY), 4.9% (DAILY)

Source: Voicebot.ai 2020
Motivation

• Search results
  • Collection of documents relevant to query, not answers

• Answers to questions
  • Specific need
  • Easily accessible to/from devices
Objective

- Tasks
- Datasets
- Models
QA Categories

• Information source
  • Text passage
  • Web documents
  • Knowledge bases
  • Tables and images
QA Categories

- Information source
- Question type
  - Factoid vs non-factoid
  - Open-domain vs closed-domain
  - Simple vs complex
Types of Questions

• Questions are very broad
  • What is 4+5?
  • How do you say [sentence] in Greek?
  • Why is ocean water salty?

• Factoid Questions
  • What is the official language of Algeria?
  • When was Marie Curie born?
    • Potentially answered using a knowledge base
QA Categories

• Information source

• Question type

• Answer type
  • Short segment
  • Paragraph
  • Yes/No
  • List
QA Tasks

• Machine reading
  • Answering questions about specific textual passages

• Open-domain question answering
  • Answering questions by collecting information from database or multiple sources
Early Work

- Simmons et al. (1964) first work on QA from text based on matching dependency parses of a question and answer
- Yale AI Project
  - Schank, Abelson, Lehnert et al. (1977)
  - Models of cognitive processes of reading comprehension
Machine Comprehension (Burges 2013)

• “A machine comprehends a passage of text if, for any question regarding that text that can be answered correctly by a majority of native speakers, that machine can provide a string which those speakers would agree both answers that question, and does not contain information irrelevant to that question.”
MCTest: A Challenge Dataset for the Open-Domain Machine Comprehension of Text

Matthew Richardson


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

2) What did James pull off of the shelves in the grocery store?
A) pudding
B) fries
C) food
D) splinters

3) Where did James go after he went to the grocery store?
A) his deck
B) his freezer
C) a fast food restaurant
D) his room
MC Test

Baselines:
1. Ngram matching: append Q &A
   Return answer with highest sentence overlap
2. Dependency parse
3. Textual Entailment

<table>
<thead>
<tr>
<th></th>
<th>MC160 Test</th>
<th>MC500 Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (SW+D)</td>
<td>66.25</td>
<td>56.67</td>
</tr>
<tr>
<td>RTE</td>
<td>59.79‡</td>
<td>53.52</td>
</tr>
<tr>
<td>Combined</td>
<td>67.60</td>
<td>60.83‡</td>
</tr>
</tbody>
</table>

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

2) What did James pull off of the shelves in the grocery store?
   A) pudding
   B) fries
   C) food
   D) splinters

Not enough training data
Kannada language is the official language of Karnataka and spoken as a native language by about 66.54% of the people as of 2011. Other linguistic minorities in the state were Urdu (10.83%), Telugu language (5.84%), Tamil language (3.45%), Marathi language (3.38%), Hindi (3.3%), Tulu language (2.61%), Konkani language (1.29%), Malayalam (1.27%) and Kodava Takk (0.18%). In 2007 the state had a birth rate of 2.2%, a death rate of 0.7%, an infant mortality rate of 5.5% and a maternal mortality rate of 0.2%. The total fertility rate was 2.2.

Q: Which linguistic minority is larger, Hindi or Malayalam?
Two Paradigms of QA (factoid)

• Knowledge-base enabled

• Information-Retrieval (IR) based
Knowledge base enabled QA

• Find semantic representation of query
  • When was Ada Lovelace born? -> birth-year (Ada Lovelace, ?x)

• Query knowledge-base
Information-Retrieval based QA

- Information-Retrieval based (aka open-domain QA)
  - Use IR to find documents with answer
  - Use machine reading comprehension to locate answer
  - Retrieve answer from a passage/document

Q: When was the premiere of *The Magic Flute*?

A: 1791
Information-Retrieval based QA

Q: When was the premiere of The Magic Flute?

A: 1791

\[ h_q = \text{BERT}_Q(q) [\text{CLS}] \]
\[ h_d = \text{BERT}_D(d) [\text{CLS}] \]
\[ \text{score}(d, q) = h_q \cdot h_d \]
IBM Watson

- Questions full of subtlety, puns and wordplay
- Clue: “Colorful fourteenth century plague that became a hit play by Arthur Miller.”
- Response: “What is The Black Death of a Salesman?”
- Exact answer not available
- Putting together pieces of information from various sources, because the exact answer is not likely to be written anywhere

Category: Diplomatic Relations
Clue: Of the four countries in the world that the United States does not have diplomatic relations with, the one that’s farthest north.
Inner subclue: The four countries in the world that the United States does not have diplomatic relations with (Bhutan, Cuba, Iran, North Korea).
Outer subclue: Of Bhutan, Cuba, Iran, and North Korea, the one that’s farthest north.
Answer: North Korea
DeepQA

Indexed several resources: Wikipedia, encyclopedias, Wordnet, semantic resources

Massively parallel: Several hypotheses and interpretations

Ferrucci et al (2010)
Figure 23.17 The 4 broad stages of Watson QA: (1) Question Processing, (2) Candidate Answer Generation, (3) Candidate Answer Scoring, and (4) Answer Merging and Confidence Scoring.
QA tasks

- Question answering: questions are in natural language
  - Answers: multiple choice, require picking from the passage, or generate freeform answer (last is pretty rare)
- Require human annotation
- “Cloze” task: word (often an entity) is removed from a sentence
  - Answers: multiple choice, pick from passage, or pick from vocabulary
  - Can be created automatically from things that aren’t questions
Datasets

• With rapid progress in deep learning
  • Renewed interest shown in multitude of datasets (30+)
  • Train supervised approaches

• Hermann et al. (NIPS 2015) DeepMind CNN/DM dataset
• Rajpurkar et al. (EMNLP 2016) SQuAD
• MS MARCO, TriviaQA, RACE, NewsQA, NarrativeQA, ...
MR as Cloze Task

"Well, Miss Maxwell, I think it only fair to tell you that you may have trouble with those boys when they do come. Forewarned is forearmed, you know. Mr. Cropper was opposed to our hiring you. Not, of course, that he had any personal objection to you, but he is set against female teachers, and when a Cropper is set there is nothing on earth can change him. He says female teachers can't keep order. He's started in with a spite at you on general principles, and the boys know it. They know he'll back them up in secret, no matter what they do, just to prove his opinions. Cropper is sly and slippery, and

Mr. Baxter privately had no hope that they would, but Esther hoped for the best. She could not believe that Mr. Cropper would carry his prejudices into a personal application. This conviction was strengthened when he overtook her walking from school the next day and drove her home. He was a big, handsome man with a very suave, polite manner. He asked interestingly about her school and her work, hoped she was getting on well, and said he had two young rascals of his own to send soon. Esther felt relieved. She thought that

?? ?? had exaggerated matters a little.

Children’s Book Test: take a section of a children’s story, block out an entity and predict it (one-doc multi-sentence cloze task)
Mr. Baxter privately had no hope that they would, but Esther hoped for the best. She could not believe that Mr. Cropper would carry his prejudices into a personal application. This conviction was strengthened when he overtook her walking from school the next day and drove her home. He was a big, handsome man with a very suave, polite manner. He asked interestedly about her school and her work, hoped she was getting on well, and said he had two young rascals of his own to send soon. Esther felt relieved. She thought that had exaggerated matters a little.

- Predict next word with LSTM LM
- Context: either just the current sentence (query) or the whole document up to this point (query+context)

Hill et al. (2015)
IR-Based QA

• Stage 1: Retrieve using IR

• Stage 2: Machine Reading
  • Currently most systems do extractive QA
    • Answer is a span of text in the passage
    • Modeled by span labeling: identifying in the passage a span that constitutes an answer
    • Given a question $q$ of $n$ tokens $q_1, \ldots, q_n$ and a passage $p$ of $m$ tokens $p_1, \ldots, p_m$ compute the probability $P(a|q, p)$ that each possible span $a$ is the answer
Beyoncé Giselle Knowles-Carter (born September 4, 1981) is an American singer, songwriter, record producer and actress. Born and raised in Houston, Texas, she performed in various singing and dancing competitions as a child, and rose to fame in the late 1990s as lead singer of R&B girl-group Destiny’s Child. Managed by her father, Mathew Knowles, the group became one of the world’s best-selling girl groups of all time. Their hiatus saw the release of Beyoncé’s debut album, Dangerously in Love (2003), which established her as a solo artist worldwide, earned five Grammy Awards and featured the Billboard Hot 100 number-one singles “Crazy in Love” and “Baby Boy”.

Q: “In what city and state did Beyoncé grow up?”
A: “Houston, Texas”

Q: “What areas did Beyoncé compete in when she was growing up?”
A: “singing and dancing”

Q: “When did Beyoncé release Dangerously in Love?”
A: “2003”
SQuAD 1.0 and 2.0

Beyoncé Giselle Knowles-Carter (born September 4, 1981) is an American singer, songwriter, record producer and actress. Born and raised in Houston, Texas, she performed in various singing and dancing competitions as a child, and rose to fame in the late 1990s as lead singer of R&B girl-group Destiny’s Child. Managed by her father, Mathew Knowles, the group became one of the world’s best-selling girl groups of all time. Their hiatus saw the release of Beyoncé’s debut album, Dangerously in Love (2003), which established her as a solo artist worldwide, earned five Grammy Awards and featured the Billboard Hot 100 number-one singles “Crazy in Love” and “Baby Boy”.

Q: “In what city and state did Beyoncé grow up?”
A: “Houston, Texas”

Q: “What areas did Beyoncé compete in when she was growing up?”
A: “singing and dancing”

Q: “When did Beyoncé release Dangerously in Love?”
A: “2003”

• Annotators given Wikipedia article
  • Create questions based on article
  • Problem?

• TriviaQA meant to address that
Neural Models for MR

- Problem formulation
  - Input: $C = (c_1, c_2, \ldots, c_N)$, $Q = (q_1, q_2, \ldots, q_M)$, $c_i, q_i \in V$
  - Output: $1 \leq \text{start} \leq \text{end} \leq N$

- A family of LSTM-based models with attention (2016–2018)
  - Attentive Reader (Hermann et al., 2015), Stanford Attentive Reader (Chen et al., 2016), Match-LSTM (Wang et al., 2017), BiDAF (Seo et al., 2017), Dynamic coattention network (Xiong et al., 2017), DrQA (Chen et al., 2017), R-Net (Wang et al., 2017), ReasoNet (Shen et al., 2017).

- Fine-tuning BERT-like models for reading comprehension (2019+)
  - Key idea: Need to model which words in the passage are most relevant to the question (and question words)
Bidirectional Attention Flow (BiDAF)

Each passage word now “knows about” the query
Bidirectional Attention Flow (BiDAF)

• **Attention Flow layer**
  • **Idea:** attention should flow both ways – from the context to the question and from the question to the context
Pretraining + Finetuning

\[ \mathcal{L} = - \log p_{\text{start}}(s^*) - \log p_{\text{end}}(e^*) \]

\[ p_{\text{start}}(i) = \text{softmax}_i(w_{\text{start}}^T h_i) \]

\[ p_{\text{end}}(i) = \text{softmax}_i(w_{\text{end}}^T h_i) \]

where \( h_i \) is the hidden vector of \( c_i \), returned by BERT

All BERT parameters (e.g., 110M) including newly introduced parameters (w’s) optimized for L

Slide credits: Danqi Chen
## Pretraining + Finetuning

**Evaluation**: exact match (0 or 1) and F1 (partial credit)

<table>
<thead>
<tr>
<th>Model</th>
<th>F1</th>
<th>EM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human performance</td>
<td>91.2*</td>
<td>82.3*</td>
</tr>
<tr>
<td>BiDAF</td>
<td>77.3</td>
<td>67.7</td>
</tr>
<tr>
<td>BERT-base</td>
<td>88.5</td>
<td>80.8</td>
</tr>
<tr>
<td>BERT-large</td>
<td>90.9</td>
<td>84.1</td>
</tr>
<tr>
<td>XLNet</td>
<td>94.5</td>
<td>89.0</td>
</tr>
<tr>
<td>RoBERTa</td>
<td>94.6</td>
<td>88.9</td>
</tr>
<tr>
<td>ALBERT</td>
<td>94.8</td>
<td>89.3</td>
</tr>
</tbody>
</table>
Pretraining + Finetuning

- For development and testing sets, 3 gold answers are collected, because there could be multiple plausible answers.
- We compare the predicted answer to each gold answer (a, an, the, punctuations are removed) and take max scores. Finally, we take the average of all the examples for both exact match and F1.
- Estimated human performance: EM = 82.3, F1 = 91.2

Q: What did Tesla do in December 1878?
A: {left Graz, left Graz, left Graz and severed all relations with his family}

Prediction: {left Graz and served}

Exact match: $\max\{0, 0, 0\} = 0$
F1: $\max\{0.67, 0.67, 0.61\} = 0.67$
Modified Pretraining Objective: SpanBERT

Two ideas:

1) masking contiguous spans of words instead of 15% random words

2) using the two end points of span to predict all the masked words in between = compressing the information of a span into its two endpoints

$$\mathcal{L}(\text{football}) = \mathcal{L}_{\text{MLM}}(\text{football}) + \mathcal{L}_{\text{SBO}}(\text{football})$$

$$= - \log P(\text{football} \mid x_7) - \log P(\text{football} \mid x_5, x_9, p_3)$$
Is QA a solved Problem?

- The current systems still perform poorly on adversarial examples or examples from out-of-domain distributions

---

**Article:** Super Bowl 50  
**Paragraph:** “Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver’s Executive Vice President of Football Operations and General Manager. *Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.*”

**Question:** “What is the name of the quarterback who was 38 in Super Bowl XXXIII?”

**Original Prediction:** John Elway  
**Prediction under adversary:** Jeff Dean

---

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>71.4</td>
<td>75.4</td>
<td>75.5</td>
<td>80.0</td>
</tr>
<tr>
<td>ADDSENT</td>
<td>27.3</td>
<td>29.4</td>
<td>34.3</td>
<td>34.2</td>
</tr>
<tr>
<td>ADDONESENT</td>
<td>39.0</td>
<td>41.8</td>
<td>45.7</td>
<td>46.9</td>
</tr>
<tr>
<td>ADDANY</td>
<td>7.6</td>
<td>11.7</td>
<td>4.8</td>
<td>2.7</td>
</tr>
<tr>
<td>ADDCOMMON</td>
<td>38.9</td>
<td>51.0</td>
<td>41.7</td>
<td>52.6</td>
</tr>
</tbody>
</table>

(Jia and Liang, 2017): Adversarial Examples for Evaluating Reading Comprehension Systems
Is QA a solved Problem?

<table>
<thead>
<tr>
<th>Fine-tuned on</th>
<th>Evaluated on</th>
<th>SQuAD</th>
<th>TriviaQA</th>
<th>NQ</th>
<th>QuAC</th>
<th>NewsQA</th>
</tr>
</thead>
<tbody>
<tr>
<td>SQuAD</td>
<td>75.6</td>
<td>46.7</td>
<td>48.7</td>
<td>20.2</td>
<td>41.1</td>
<td></td>
</tr>
<tr>
<td>TriviaQA</td>
<td>49.8</td>
<td>58.7</td>
<td>42.1</td>
<td>20.4</td>
<td>10.5</td>
<td></td>
</tr>
<tr>
<td>NQ</td>
<td>53.5</td>
<td>46.3</td>
<td>73.5</td>
<td>21.6</td>
<td>24.7</td>
<td></td>
</tr>
<tr>
<td>QuAC</td>
<td>39.4</td>
<td>33.1</td>
<td>33.8</td>
<td>33.3</td>
<td>13.8</td>
<td></td>
</tr>
<tr>
<td>NewsQA</td>
<td>52.1</td>
<td>38.4</td>
<td>41.7</td>
<td>20.4</td>
<td>60.1</td>
<td></td>
</tr>
</tbody>
</table>
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!)
  - And, remember it (SQuAD 2.0)
Open-Domain QA

- Contrast to closed-domain systems (medicine, technical support)

- Don’t assume a passage is given; instead, consider a collection of documents
How many of Warsaw's inhabitants spoke Polish in 1933?

Retriever-Reader

https://github.com/facebookresearch/DrQA
Information-Retrieval based QA

- Each text passage can be encoded as a vector using BERT and the retriever score can be measured as the dot product between the question representation and passage representation.

- However, it is not easy to model as there are a huge number of passages (e.g., 21M in English Wikipedia)

Lee et al., 2019. Latent Retrieval for Weakly Supervised Open Domain Question Answering

Slide credits: Danqi Chen
Retrieval is Trainable

- Dense passage retrieval (DPR) - We can also just train the retriever using question-answer pairs!

\[
sim(q, p) = h_q^T h_p
\]

- Trainable retriever (using BERT) largely outperforms traditional IR retrieval models

Karpukhin et al., 2020. Dense Passage Retrieval for Open-Domain Question Answering

Slide credits: Danqi Chen
Large Language Models do QA

- ... without an explicit retriever stage

Roberts et al., 2020. How Much Knowledge Can You Pack Into the Parameters of a Language Model?
Takeaways

• Many flavors of reading comprehension tasks: cloze or actual questions, single or multi-sentence

• Complex attention schemes can match queries against input texts and identify answers

• Multi-hop Question Answering - ability to answer questions that draw on several sentences or several documents to answer
Limited QA

• Focus on questions such as test answers
  • *What were the main causes of World War 1?*
  • Summarization (multi-document?)
  • *Can you get the flu from a flu shot?*
  • Need an explanation, not just Yes/No
  • *What temperature should I heat the milk to?*
  • Potentially listed in KB but not really
Beyond Textual QA

Visual QA [Antol et al., 2015]