Natural Language Processing

Question Answering
Dan Klein – UC Berkeley

The following slides are largely from Greg Durrett and Chris Manning, including many slides originally from Sanda Harabagiu, ISI, and Nicholas Kushmerick.
QA is Very Broad

- Factoid QA: *what states border Mississippi?, when was Barack Obama born?*
  - Lots of this could be handled by QA from a knowledge base, if we had a big enough knowledge base
- “Question answering” as a term is so broad as to be meaningless
  - *What is the meaning of life?*
  - *What is 4+5?*
  - *What is the translation of [sentence] into French?* [McCann et al., 2018]
QA Limits

- Focus on questions where the answer might appear in text — still hard!
  - *What were the main causes of World War II?* — requires summarization

- *Can you get the flu from a flu shot?* — want IR to provide an explanation of the answer, not just yes/no

- *How long should I soak dry pinto beans?* — could be written down in a KB but probably isn’t

- Today: QA when it requires retrieving the answer from a passage
Related: Reading Comprehension

- “AI challenge problem”: answer question given context
- Recognizing Textual Entailment (2006)
- MCTest (2013): 500 passages, 4 questions per passage
- Two questions per passage explicitly require cross-sentence reasoning

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.

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

Richardson (2013)
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 went home.

Where did James go after he went to the grocery store?

James went to the fast food restaurant after the grocery store.
A Brief (Academic) History

- Question answering is not a new research area
- Question answering systems can be found in many areas of NLP:
  - Natural language / database systems
    - A lot of early NLP work on these
  - Conversational / assistant systems
    - Currently very active and commercially relevant
- The focus on open-domain QA is (relatively) new
  - TREC QA competition: 1999+
  - Modern large-scale factoid QA, eg SQuAD: 2016+
  - Search increasingly includes question answering
- General approach (across all eras): retrieval + entailment
Classic Question Answering

- Form semantic representation from semantic parsing, execute against structured knowledge base

Q: *where was Barack Obama born*

\[ \lambda x. \text{type}(x, \text{Location}) \land \text{born_in}(\text{Barack_Obama}, x) \]

(also Prolog / GeoQuery, etc.)

- How to deal with open-domain data/relations? Need data to learn how to ground every predicate or need to be able to produce predicates in a zero-shot way
Famous QA Example: Watson (2011)
Category: General Science
Clue: When hit by electrons, a phosphor gives off electromagnetic energy in this form.
Answer: Light (or Photons)

Category: Lincoln Blogs
Clue: Secretary Chase just submitted this to me for the third time; guess what, pal. This time I’m accepting it.
Answer: his resignation

Category: Head North
Clue: They’re the two states you could be reentering if you’re crossing Florida’s northern border.
Answer: Georgia and Alabama

Category: Decorating
Clue: Though it sounds “harsh,” it’s just embroidery, often in a floral pattern, done with yarn on cotton cloth.
Answer: crewel

Category: “Rap” Sheet
Clue: This archaic term for a mischievous or annoying child can also mean a rogue or scamp.
Subclue 1: This archaic term for a mischievous or annoying child.
Subclue 2: This term can also mean a rogue or scamp.
Answer: Rapscallion

Category: Before and After Goes to the Movies
Clue: Film of a typical day in the life of the Beatles, which includes running from bloodthirsty zombie fans in a Romero classic.
Subclue 2: Film of a typical day in the life of the Beatles.
Answer 1: (A Hard Day’s Night)
Answer 2: (Night of the Living Dead)
Answer: A Hard Day’s Night of the Living Dead
Architecture
Retrieval
Querying Documents with Keywords

- **Goal:** take a query and find relevant documents
- **Example:** songs in a database
- **Constraints:** want documents that contain query words, but not exact match on query...
- **Idea:** rank documents by how much of the query they contain?

**Query:**
we wait on trains

**New Romantics**

- *Fifteen*
  - We're all bored
  - We're all so tired of everything
  - We wait for trains that just aren't coming

- *We show off our different scarlet letters*
  - Trust me, mine is better

- *But*...

- *He*
  - Drive
  - I think

- *You*
  - And it's
  - Not away

- *But*...
Term Frequency

- **Idea:** Score of document by summing over query words:

\[
score(q, d) = \sum_{w \in q} score(w, d) \quad \text{where} \quad score(w, d) = (w \in d)
\]

- **Problem:** Many documents could contain all query words (e.g., once)
- **Solution:** Term frequency: \( score(q_i, d) = \text{tf}(w, d) := C(w, d) \)

<table>
<thead>
<tr>
<th></th>
<th>Out of the Woods</th>
<th>Speak Now</th>
<th>Fifteen</th>
<th>New Romantics</th>
<th>Tim McGraw</th>
<th>CMUDICT</th>
</tr>
</thead>
<tbody>
<tr>
<td>“we”</td>
<td>73</td>
<td>2</td>
<td>12</td>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>“wait”</td>
<td>4</td>
<td>1</td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>“on”</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>“trains”</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>
Saturation and Normalization

- Problem: One query term being repeated many times can outweigh other terms being entirely absent
- Solution: Give counts of any one term diminishing returns
- Example: (Log) Normalization: \( \text{tf}(w, d) := \log(C(w, d) + 1) \)

<table>
<thead>
<tr>
<th></th>
<th>Out of the Woods</th>
<th>Speak Now</th>
<th>Fifteen</th>
<th>New Romantics</th>
<th>Tim McGraw</th>
</tr>
</thead>
<tbody>
<tr>
<td>“we”</td>
<td>4.30</td>
<td></td>
<td>1.01</td>
<td>2.56</td>
<td>1.39</td>
</tr>
<tr>
<td>“wait”</td>
<td></td>
<td>1.61</td>
<td></td>
<td>0.69</td>
<td></td>
</tr>
<tr>
<td>“on”</td>
<td>0.69</td>
<td>1.39</td>
<td>1.01</td>
<td>1.61</td>
<td>2.40</td>
</tr>
<tr>
<td>“trains”</td>
<td></td>
<td></td>
<td></td>
<td>0.69</td>
<td></td>
</tr>
</tbody>
</table>
Weighting Terms with TF-IDF

- **Problem:** Common words will have high counts but carry little information

- **Solution:** Downweight words that occur in many documents

- **Inverse document frequency (IDF):**
  - Basic document frequency of $w$ is fraction of documents with the $w$: $\frac{D(w)}{N}$

- **TF-IDF: Classic IR baseline**
  - Normalized term frequency: $\text{tf}(w, d) := \log(C(w, d) + 1)$
  - Normalized inverse document frequency: $\text{idf}(w) := \log \left( \frac{N}{D(w)} \right)$
  - $\text{TF-IDF}(w, d) = \text{tf}(w, d) \times \text{idf}(w)$
Weighting Terms with TF-IDF

- Term frequency: \( tf(w, d) := \log(C(w, d) + 1) \)
- Inverse document frequency: \( idf(w) := \log \left( \frac{N}{D(w)} \right) \)
- \( TF-IDF(w, d) = tf(w, d) \times idf(w) \)

<table>
<thead>
<tr>
<th></th>
<th>Out of the Woods</th>
<th>Speak Now</th>
<th>Fifteen</th>
<th>New Romantics</th>
<th>Tim McGraw</th>
</tr>
</thead>
<tbody>
<tr>
<td>“we”</td>
<td>4.30</td>
<td>1.01</td>
<td>2.56</td>
<td>1.39</td>
<td></td>
</tr>
<tr>
<td>“wait”</td>
<td>1.61</td>
<td></td>
<td>0.69</td>
<td></td>
<td></td>
</tr>
<tr>
<td>“on”</td>
<td>0.69</td>
<td>1.39</td>
<td>1.01</td>
<td>1.61</td>
<td>2.40</td>
</tr>
<tr>
<td>“trains”</td>
<td>0.69</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Weighting Terms with TF-IDF

- Term frequency: $tf(w, d) := \log(C(w, d) + 1)$
- Inverse document frequency: $idf(w) := \log\left(\frac{N}{D(w)}\right)$
- $TF-IDF(w, d) = tf(w, d) \times idf(w)$

<table>
<thead>
<tr>
<th>Term</th>
<th>Out of the Woods</th>
<th>Speak Now</th>
<th>Fifteen</th>
<th>New Romantics</th>
<th>Tim McGraw</th>
<th>IDF</th>
</tr>
</thead>
<tbody>
<tr>
<td>“we”</td>
<td>2.47</td>
<td></td>
<td>0.63</td>
<td>1.47</td>
<td>0.79</td>
<td>0.57</td>
</tr>
<tr>
<td>“wait”</td>
<td></td>
<td>3.78</td>
<td></td>
<td>1.63</td>
<td></td>
<td>2.35</td>
</tr>
<tr>
<td>“on”</td>
<td>0.22</td>
<td>0.45</td>
<td>0.36</td>
<td>0.52</td>
<td>0.78</td>
<td>0.32</td>
</tr>
<tr>
<td>“trains”</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4.54</td>
</tr>
<tr>
<td>Sum of terms</td>
<td>2.69</td>
<td>4.22</td>
<td>0.99</td>
<td><strong>6.77</strong></td>
<td>1.57</td>
<td></td>
</tr>
</tbody>
</table>
BM25: Addressing Issues with TF-IDF

Some remaining issues with TF-IDF:

- Handling of term saturation
- Varying document length

Given a query $w = \{w_1, w_2, \ldots, w_n\}$:

$$BM25(w, d) := \sum_{i=1}^{n} \text{idf}(w_i) \cdot \frac{\text{tf}(w_i, d) \cdot (k + 1)}{\text{tf}(w_i, d) + (1 - b + b \cdot \frac{|d|}{L})}$$

Still a strong zero-shot retrieval baseline as of December 2021
Beyond Term Scoring

- Term scoring suited classic inverted index construction very well!

- Modern IR systems include other term factors
  - Contiguous match (e.g. n-grams)
  - Positional information (e.g. titles)
  - Related word match (e.g. synonyms)

- ... And some of the most important features aren’t term-derived at all
  - Link analysis (e.g. PageRank)
  - User behavior (e.g. clickstream analysis)
Neural Retrieval
Dense Passage Retrieval

Contrastive loss function:

$$- \log \frac{\exp(\text{sim}(w_i, d_i^+))}{\exp(\text{sim}(w_i, d_i^+)) + \sum_{j=1}^{n} \exp(\text{sim}(w_i, d_{i,j}^-))}$$

where $d_{i,j}^-$ are negatives and $\text{sim}$ is vector similarity.

Obtain “hard negatives” using incorrect answers from a BM25 baseline model

[Karpukhin et al, 2020]
Retrieval-Augmented Models

- Generation models (including QA) have a bottleneck where parameters must capture all information from the training data.
- Retrieval-augmented models let a system look directly at source data.
- Similar to how attention let encoder-decoders look directly at the input
- Also allows a system to dynamically respond to new data after training

[Lewis et al, 2020]
Nearest Neighbor Language Modeling

- Nonparametrics: condition generation on retrieved documents
- Slightly improves perplexity at the cost of inference speed and storage, but can update knowledge without retraining

[Image from Khandelwal, et al. (2019): Nearest Neighbor Language Models]
Question Answering
Web Question Answering: Is More Always Better?

[Dumais, Banko, Brill, Lin, Ng 2002]

Q: “Where is the Louvre located?”

Want “Paris” or “France”, etc

These answers are often all over the documents returned by a web search

Idea: the answer is probably a frequent n-gram in the search results
AskMSR: Shallow approach

- In what year did Abraham Lincoln die?
- Ignore hard documents and find easy ones
AskMSR: Details

1. Question
   "Where is the Louvre Museum located?"

   in Paris France 59%
   museums 12%
   hostels 10%

2. Rewrite Query

3. <Search Engine>

4. Collect Summaries, Mine N-grams

5. Tile N-Grams

4. Filter N-Grams

N-Best Answers
Results

- Standard TREC contest test-bed:
  ~1M documents; 900 questions

- Technique doesn’t do too well (though would have placed in top 9 of ~30 participants!)
  - MRR = 0.262 (ie, right answered ranked about #4-#5 on average)
  - Why? Because it relies on the redundancy of the Web

- Using the Web as a whole, not just TREC’s 1M documents...
  MRR = 0.42 (ie, on average, right answer is ranked about #2-#3)
Abduction: LCC
LCC: Harabagiu, Moldovan et al.
Abductive inference

- **System attempts inference to justify an answer (often following lexical chains)**
  - This inference is a kind of middle ground between logic and pattern matching
  - ... but it can be effective: 30% improvement at the time

- **Example:**
  - **Q:** *When was the internal combustion engine invented?*
  - **A:** *The first internal-combustion engine was built in 1867.*
  - *invent -> create_mentally -> create -> build*
Question Answering Example

- How hot does the inside of an active volcano get?

- “lava fragments belched out of the mountain were as hot as 300 degrees Fahrenheit”
  - volcano ISA mountain
  - lava ISPARTOF volcano ■ lava IN volcano
  - fragments of lava HAVEPROPERTIESOF lava

- The needed semantic information is in WordNet definitions, and was successfully translated into a form that was used for rough ‘proofs’
Span-Based QA
SQuAD

- Single-document question-answering task where the answer is always a substring of the passage (= a paragraph from Wikipedia)
- Predict start and end indices of the answer in the passage

Rajpurkar et al. (2016)
Just Seq2Seq?

What was Marie Curie the first female recipient of?

- Like a tagging problem over the sentence (not multiclass classification), but we need some way of attending to the query

Rajpurkar et al. (2016)
A Simple Neural Architecture

- Predict a distributions over start and end points of the answer

\[ P(\text{end} \mid q, p) \] computed similarly

\[ P(\text{start} = i \mid q, p) = \text{softmax}(p_i^T W q) \]

encoding of passage

BiLSTM encoder

Marie Curie was the first ...

Who was the first female recipient of the Nobel Prize?
Training

- Train on labeled data with start and end points, maximize likelihood of correct decisions:
  \[
  \log \sum_{i \in \text{gold starts}} p(\text{start} = i | p, q) + \log \sum_{i \in \text{gold ends}} p(\text{end} = i | p, q)
  \]

In September 1958, Bank of America launched a new product called BankAmericard in Fresno. After a troubled gestation during which its creator resigned, BankAmericard went on to become the first successful credit card; that is, a financial instrument that was usable across a large number of merchants and also allowed card holders to revolve a balance (earlier financial products could do one or the other but not both). In 1976, BankAmericard was renamed and spun off into a separate company known today as Visa Inc.

What was the name of the first successful credit card?

- Inference: maximize P(start) + P(end) with the constraint that (start, end) isn’t too big a span
Some Outputs

**Question:** who caught a 16-yard pass on this drive?

**Answer:** devin funchess

there would be no more scoring in the third quarter, but early in the fourth, the broncos drove to the panthers 41-yard line. on the next play, ealy knocked the ball out of manning 's hand as he was winding up for a pass, and then recovered it for carolina on the 50-yard line. a 16-yard reception by devin funchess and a 12-yard run by stewart then set up gano 's 39-yard field goal, cutting the panthers deficit to one score at 16â€“10. the next three drives of the game would end in punts.
Question: how many victorians are non-religious?

Answer: 20%

About 61.1% of Victorians describe themselves as Christian. Roman Catholics form the single largest religious group in the state with 26.7% of the Victorian population, followed by Anglicans and members of the Uniting Church. Buddhism is the state’s largest non-Christian religion, with 168,637 members as of the most recent census. Victoria is also home to 152,775 Muslims and 45,150 Jews. Hinduism is the fastest growing religion. Around 20% of Victorians claim no religion. Amongst those who declare a religious affiliation, church attendance is low.

About 61.1% of Victorians describe themselves as Christian. Roman Catholics form the single largest religious group in the state with 26.7% of the Victorian population, followed by Anglicans and members of the Uniting Church. Buddhism is the state’s largest non-Christian religion, with 168,637 members as of the most recent census. Victoria is also home to 152,775 Muslims and 45,150 Jews. Hinduism is the fastest growing religion. Around 20% of Victorians claim no religion. Amongst those who declare a religious affiliation, church attendance is low.
Why did SQuAD Take Off?

- SQuAD was **big**: >100,000 questions at a time when deep learning was exploding

- SQuAD was **pretty easy**: year-over-year progress for a few years until the dataset was essentially solved

- SQuAD had **room to improve**: ~50% performance from a logistic regression baseline (classifier with 180M features over constituents)
Example: Richer Model Structures

Each passage word now “knows about” the query

Seo et al. (2016)
Pre-Training

What was Marie Curie the first female recipient of? [SEP] Marie Curie was the first female recipient of ...

- Predict start and end positions in passage
- No need for cross-attention mechanisms!

Devlin et al. (2019)
# Leaderboards

## 2018

<table>
<thead>
<tr>
<th>Rank</th>
<th>Model</th>
<th>EM</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>BERT (ensemble)</td>
<td>87.433</td>
<td>93.160</td>
</tr>
<tr>
<td></td>
<td>Google AI Language</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>BERT (single model)</td>
<td>85.083</td>
<td>91.835</td>
</tr>
<tr>
<td></td>
<td>Google AI Language</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>nlnet (ensemble)</td>
<td>85.356</td>
<td>91.202</td>
</tr>
<tr>
<td></td>
<td>Microsoft Research Asia</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>nlnet (ensemble)</td>
<td>85.954</td>
<td>91.677</td>
</tr>
<tr>
<td></td>
<td>Microsoft Research Asia</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>QANet (ensemble)</td>
<td>84.454</td>
<td>90.490</td>
</tr>
<tr>
<td></td>
<td>Google Brain &amp; CMU</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>r-net (ensemble)</td>
<td>84.003</td>
<td>90.147</td>
</tr>
<tr>
<td></td>
<td>Microsoft Research Asia</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

## 2021

<table>
<thead>
<tr>
<th>Rank</th>
<th>Model</th>
<th>EM</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>FPNet (ensemble)</td>
<td>90.871</td>
<td>93.183</td>
</tr>
<tr>
<td></td>
<td>Ant Service Intelligence Team</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>IE-Net (ensemble)</td>
<td>90.758</td>
<td>93.044</td>
</tr>
<tr>
<td></td>
<td><a href="https://arxiv.org/abs/1810.04805">RICOH_SRCP_DML</a></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>SA-Net on Albert (ensemble)</td>
<td>90.724</td>
<td>93.011</td>
</tr>
<tr>
<td></td>
<td>QIANXIN</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>SA-Net-V2 (ensemble)</td>
<td>90.679</td>
<td>92.948</td>
</tr>
<tr>
<td></td>
<td>QIANXIN</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Retro-Reader (ensemble)</td>
<td>90.578</td>
<td>92.978</td>
</tr>
<tr>
<td></td>
<td>Shanghai Jiao Tong University</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>FPNet (ensemble)</td>
<td>90.600</td>
<td>92.899</td>
</tr>
<tr>
<td></td>
<td>YuYang</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>TransNets + SFVerifier + SFEEnsembler (ensemble)</td>
<td>90.487</td>
<td>92.894</td>
</tr>
<tr>
<td></td>
<td>Senseforth AI Research</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><a href="https://www.senseforth.ai/">https://www.senseforth.ai/</a></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>