Question Answering

CMSC 473/673 Spring 2017
Bryan Wilkinson
The Alaska pipeline starts at the frozen edge of the Arctic Ocean. It stretches southward across the largest and northernmost state in the United States, ending at a remote ice-free seaport village nearly 800 miles from where it begins. It is massive in size and extremely complicated to operate. The steel pipe crosses windswept plains and endless miles of delicate tundra that tops the frozen ground. It weaves through crooked canyons, climbs sheer mountains, plunges over rocky crags, makes its way through thick forests, and passes over or under hundreds of rivers and streams. The pipe is 4 feet in diameter, and up to 2 million barrels (or 84 million gallons) of crude oil can be pumped through it daily.

According to the passage, 84 million gallons of oil can travel through the pipeline each
A. day         B. week        C. month          D. year

Example taken from TOEFL practice test.
Single Document Question Answering

- Also known as comprehension tests in the literatures
- Numerous datasets have been proposed in the past year or so
  - Machine Comprehension Test (2013)
  - Stanford Question Answering Dataset (SQUAD) 2016
- Most approaches to this task use Neural Networks and Distributional Semantics
  - Passages are usually short, and therefore cheap to encode
  - No searching needed
Multi Document Question Answering

- Find the answer to a question in a large corpus
  - The Web
  - Wikipedia
- OR in a pre-constructed knowledge base

A Big Corpus

84 million gallons of oil can travel through the Alaska pipeline each?

A. day    B. week    C. month    D. year
Question Answering Process

- In all cases Question Answering (QA) can be broken down into 3 parts
  - The methods and details for each part will be different depending on the situation

1. Formulate a query
   - Take what the user has asked (or the test) and turn it into something that can be used to help find the answer

2. Find the answer
   - Search through documents, sentences, or a database, to find the text that contains the answer

3. Formulate the answer
   - Extract the answer from the text and possibly generate a new sentence containing that answer.

84 million gallons of oil can travel through the Alaska pipeline each? 
A. day         B. week        C. month          D. year
Information Retrieval deals with how we store and retrieve information.

It is a large field with many complex methods:
- Modern methods take into account that the document space will be continuously expanding.
- Also based on more than just the words in the text.

Today we will just look at a simple vector based approach so we have something to base the rest of the discussion on.
- We’ve actually already used it in summarization.
Each document is represented as a bag of words
- No position information is retained, only how many times a given word appeared in each document

Since we are dealing with counts, we represent each document as a vector of counts
- Normally very common words, known as stopwords, are not counted

In order to determine which words are more important to which documents we weight the cells of the matrix using TF-IDF

Document similarity is the cosine similarity between two documents
Sidebar: Information Retrieval Libraries

- The Vector Space model is a good basis, but is very simple, and says nothing about storage and retrieval.
- Many published papers rely on web search engines for retrieval but this is also not ideal.
- If you were to implement your own QA system, an IR library becomes very useful:
  - They are optimized for quick storage and retrieval.
  - They can store a document with many attributes.
  - They have built in pre-processing, like stemming, stopwords, etc.
- Lucene (for Java) is probably the most popular, but many others exist.
Query Formation (From Question to Query)

- We need to form a vector from the question, there are several ways to do that.
- The simplest is to take all words in the question and set the corresponding dimensions in the vector to 1.

Who wrote the Magic School Bus?

<table>
<thead>
<tr>
<th></th>
<th>Bus</th>
<th></th>
<th>Magic</th>
<th></th>
<th>School</th>
<th></th>
<th>the</th>
<th>who</th>
<th></th>
<th>wrote</th>
</tr>
</thead>
<tbody>
<tr>
<td>..0..</td>
<td>1</td>
<td>..0..</td>
<td>1</td>
<td>..0..</td>
<td>1</td>
<td>..0..</td>
<td>1</td>
<td>1</td>
<td>..0..</td>
<td>1</td>
</tr>
</tbody>
</table>
Who wrote the Magic School Bus?

<table>
<thead>
<tr>
<th></th>
<th>Bus</th>
<th>Magic</th>
<th>School</th>
<th>the</th>
<th>who</th>
<th></th>
<th>wrote</th>
</tr>
</thead>
<tbody>
<tr>
<td>.0..</td>
<td>1</td>
<td>.0..</td>
<td>1</td>
<td>.0..</td>
<td>1</td>
<td>1</td>
<td>..0..</td>
</tr>
</tbody>
</table>

The Magic School Bus was written by Joanna Cole.

<table>
<thead>
<tr>
<th></th>
<th>Bus</th>
<th>Magic</th>
<th>School</th>
<th>the</th>
<th>who</th>
<th></th>
<th>wrote</th>
</tr>
</thead>
<tbody>
<tr>
<td>.0..</td>
<td>1</td>
<td>.0..</td>
<td>1</td>
<td>.0..</td>
<td>0</td>
<td>0</td>
<td>..0..</td>
</tr>
</tbody>
</table>

Africa is sung by David Paich, who wrote most of the album Toto IV

<table>
<thead>
<tr>
<th></th>
<th>Bus</th>
<th>Magic</th>
<th>School</th>
<th>the</th>
<th>who</th>
<th></th>
<th>wrote</th>
</tr>
</thead>
<tbody>
<tr>
<td>.0..</td>
<td>0</td>
<td>.0..</td>
<td>0</td>
<td>.0..</td>
<td>1</td>
<td>1</td>
<td>..0..</td>
</tr>
</tbody>
</table>
Additions to the Query Formulation

- Remove stop words that appear in almost every document
  - Who wrote the Magic School Bus → {Who, wrote, Magic, School, Bus}

- Remove question words - they don’t normally occur in text with answers
  - Who wrote the Magic School Bus → {wrote, Magic, School, Bus}

- Stem all words in the query AND corpus
  - Who wrote the Magic School Bus → {write, Magic, School, Bus}

- Add in synonyms - from a source like WordNet or using vector semantics
  - Who wrote the Magic School Bus → {write, create, invent, Magic, mysterious, School, Bus}
The keywords {write, create, invent, Magic, mysterious, School, Bus} return passages like

Joanna Cole (born August 11, 1944), is an American writer of children’s books. She is most famous as the author of "The Magic School Bus" series. She has written over 250 books ranging from her first book "Cockroaches" to her famous series "Magic School Bus", which is illustrated by Bruce Degen.

He may be known best for illustrating the "The Magic School Bus" series written by Joanna Cole. He has collaborated with writers Nancy White Carlstrom, on the Jesse Bear books, and Jane Yolen, on the Commander Toad series. Degen has authored "Jamberry", "Daddy Is a Doodlebug", and "Shirley's Wonderful Baby".

Gil Filar (born September 1, 1986) is a Canadian writer, ex-actor and -magician. He is known for his voicing Fox on the first three seasons of "Franklin". He is also the brother of Maia Filar, who voiced Pheobe in "The Magic School Bus".
The Magic School Bus is a series of children's books about science, written by Joanna Cole and illustrated by Bruce Degen. They feature the antics of Ms. Valerie Frizzle and her class, who board an anthropomorphic Type A school bus which takes them on field trips to impossible locations, such as the solar system, clouds, the past, and the human body. The books are written in the first person from the point of view of an unnamed student in "the Friz's" class. The class pet Liz, a lizard, accompanies the class on their field trips.

The show's theme song called "Ride on the Magic School Bus" was written by Peter Lurye and performed by rock 'n' roll legend Little Richard.
Rewriting the Query

- Sometimes standard queries evoke common words in the document
- We can exploit this by rewriting the query according to a few hand-written rules
  - Where is the Valley of the Kings → The Valley of the Kings is located in
  - When was the telephone invented → the telephone was invented
    - No change in vector space IR
  - What did [ENTITY] verb [ENTITY] → [ENTITY] verb+ed [ENTITY]
    - Doesn’t change query if we just end up stemming it
Question Classification

- Another piece of useful information the query may provide is what type of thing the answer is expected to be
  - This won’t help with retrieval, but will be very helpful in picking the final answer eventually
- Question types are hierarchical
  - The answer to the question “Which city is the capital of Maryland” is a city, but it is also a location
- Both rule-based and supervised learning approaches have been used for this task
Rule-Based Question Classification

- One example is Hovy, et al. 2002. A question/answer typology with surface text patterns.
  - The taxonomy is still available

- Name
  - What ADJ family ...
  - What is/was the surname/family name/last name of X

- City
  - What is the capital of PROPER-PLACE
  - What is/are (the SUPERLATIVE) town/city/hub/urban area
Supervised Learning Question Classification

- Rather than building rules by hand, a common approach is to train a classifier with a question as input, and the expected answer as the label.

- **Common Features**
  - Words in the question
  - Part of speeches in the question
  - Named entities in the question
  - Named entities types in the question

- **One very useful feature is the headword of the first NP after the question word**
  - Requires parsing or at least chunking
  - Example: What is the state **flower** of California
  - Example: Which **city** in China has the largest number of companies.

Examples modified from SLP2
For our question:

Who wrote the Magic School Bus?

We can determine that the answer should be a person because the question matches the following pattern

**who <complex action***>

Which is a rule that requires a person answer according to Hovy’s taxonomy.

***“complex actions can have many variations and paraphrases. Here: {set foot, step, walk, stand} on the moon, or “touch the surface of the moon””***
Finding the Answer in the Returned Documents

• We can think of the set of documents returned by the query as a new smaller corpora to work with
• It is much easier to do more intense processing on this smaller set of documents
• Finding the answer, or at least more closely finding the answer, in this reduced set is called passage retrieval
  ○ The passage might be a sentence or a paragraph
  ○ Could be a snippet of text returned to us automatically by a search engine
• The passages need to be ranked
Ranking the Passages

- Use a small set of features to rank the passages
  - Can either build a supervised model to rank them
  - Use hand written rules about which features have priority

- Common Features
  - Number of Named Entities of correct type (IE person, place, animal, etc)
  - Number of Keywords in Passage
  - Longest sequence match to query in passage
  - Proximity of query words in passage (e.g. Are they all within a 5 word window?)
  - N-gram overlap (Like ROUGE score) of ORIGINAL question and passage
Joanna Cole (born August 11, 1944), is an American writer of children’s books. She is most famous as the author of "The Magic School Bus" series. She has written over 250 books ranging from her first book "Cockroaches" to her famous series "Magic School Bus", which is illustrated by Bruce Degen.

# of Named Entities of Correct type (PERSON): 2
# of question keywords in passage: 4 or 7
Longest Sequence from Query: 3
Proximity of Query Words: 16 words
Bi-Gram Overlap: 2
He may be known best for illustrating the "The Magic School Bus" series written by Joanna Cole. He has collaborated with writers Nancy White Carlstrom, on the Jesse Bear books, and Jane Yolen, on the Commander Toad series. Degen has authored "Jamberry", "Daddy Is a Doodlebug", and "Shirley's Wonderful Baby".

# of Named Entities of Correct type (PERSON): 4
# of question keywords in passage: 4
Longest Sequence from Query: 3
Proximity of Query Words: 9 words
Bi-Gram Overlap: 2
Gil Filar (born September 1, 1986) is a Canadian **writer**, ex-actor and -magician. He is known for his voicing Fox on the first three seasons of "Franklin". He is also the brother of Maia Filar, who voiced Pheobe in "The Magic School Bus".

# of Named Entities of Correct type (PERSON): 2

# of question keywords in passage: 4

Longest Sequence from Query: 3

Proximity of Query Words: 34 words

Bi-Gram Overlap: 2
If we continue on with this analysis for the other two passages, we get a feature vector that looks like this.

Let’s define the score, $S$, of each passage as the sum of all features except proximity, minus proximity.

<table>
<thead>
<tr>
<th></th>
<th>#NE</th>
<th>#Keywords</th>
<th>LongSeq</th>
<th>Proximity</th>
<th>Overlap</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P_1$</td>
<td>2</td>
<td>7</td>
<td>3</td>
<td>16</td>
<td>2</td>
<td>-2</td>
</tr>
<tr>
<td>$P_2$</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>9</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>$P_3$</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>32</td>
<td>2</td>
<td>-21</td>
</tr>
<tr>
<td>$P_4$</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>15</td>
<td>2</td>
<td>-2</td>
</tr>
<tr>
<td>$P_5$</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>8</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>
According to our scoring formula, the top 3 passages are (assume we break ties of the first feature)

He may be known best for illustrating the "The Magic School Bus" series written by Joanna Cole. He has collaborated with writers Nancy White Carlstrom, on the Jesse Bear books, and Jane Yolen, on the Commander Toad series. Degen has authored "Jamberry", "Daddy Is a Doodlebug", and "Shirley's Wonderful Baby".

The show's theme song called "Ride on the Magic School Bus" was written by Peter Lurye and performed by rock 'n' roll legend Little Richard.

The Magic School Bus is a series of children's books about science, written by Joanna Cole and illustrated by Bruce Degen. They feature the antics of Ms. Valerie Frizzle and her class, who board an anthropomorphic Type A school bus which takes them on field trips to impossible locations, such as the solar system, clouds, the past, and the ....
Once we have a ranked set of passages, we need to search through them to find the answer.

Two primary ways this is done:

- **Rule Based methods**
  - Handcrafted or learned rules, based on the expected answer type
  - Multiple candidate answers are usually extracted and then ranked again

- **N-Gram Tiling**
  - First weight unigrams, bigrams, and trigrams based on the number of snippets (or passages, etc) they occur in
  - Create an answer by overlapping the highest scoring n-grams
Rule-Based Answer Extraction

- Each type of expected answer has different answer patterns associated with it.
- A pattern based on looking for a PERSON answer might be **VERB by PERSON**
- These patterns are used to extract the answers from the text, but are not normally used to pick the correct answer.
- To choose the correct answer, we construct a feature matrix for each answer candidate, and then rank as we did with passage ranking.
- Distance from answer to keyword and position in the sentence are common features.
If we apply the pattern **VERB by PERSON** to our top three passages, we get the following candidate answers:

- written by **Joanna Cole**
- was written by **Peter Lurye**
- written by **Joanna Cole**
- illustrated by **Bruce Degen**

Let our ranking function be based on distance to the keywords and if there is a punctuation after the person:

- This handcrafted rule, selects the right answer, **Joanna Cole**!
N-Gram Tiling

- N-Gram Tiling uses the intuition that the correct answer should appear in more documents
  - Collect all n-grams up to n=3, and weight them based on the number of passages they are in
  - We can additionally weight the n-grams by if they were returned with the original query or an expanded one
- Before checking for the most frequent N-grams, it is a good idea to filter them by expected answer type
- After filtering, iteratively overlap n-grams, combining the scores of the overlapped ngrams
  - At each iteration remove the lowest scored ngram
  - Repeat until only one answer is left.
From our three passages, we get the following n-grams, after filtering:

Joanna, Cole, Joanna Cole, Joanna Cole, Bruce, Degen, Bruce Degen, by Bruce, illustrated by Bruce, by Bruce Degen, Shirley, by Joanna, writers Nancy, Nancy White, White Carlstrom, and Jane, Jane Yolen, by Joanna Cole, with writers Nancy, writers Nancy White, Nancy White Carlstrom, and Jane Yolen, and Shirley's, Shirley's wonderful, written by Joanna, Peter, Lurve, Richard, Peter Lurve, Little Richard, by Peter, written by Peter, by Peter Lurve, legend Little Richard.

All have a weight of 1, except Joanna, Cole, Joanna Cole, which have a weight of 2.
N-Gram Tiling Exercise

- Greedily search for overlaps, starting with highest weighted n-gram
- Overlap Joanna and Joanna Cole, the new Joanna Cole gets a score of 4, and we delete the lowest scoring n-gram from the set.
- Overlap Joanna Cole with Joanna Cole (the new overlap) to get score of 6.
- Continue this process and you will eventually end up with all Joanna, Joanna Cole, or Cole in the dataset, with nothing left
  - At this point you will start deleting Joanna and Cole, and end up with the final answer.
Traditionally just giving the user the answer, maybe in context of a paragraph was enough.

As we move towards better human interaction with NLP systems, users may prefer to have the answer in the form of a sentence.

- We’ll talk about dialogue next class.
- A simple method we can use for now is based on patterns.

Goal:

- Given the question “How many people live in the United States?”
- Go from “319 million” to “319 people live in the United States” or even “The population of the United States is 319 million people”
Answer Template Examples

- For a question matching the pattern

  Who VERB-PAST CREATIVE-WORK

- We might have the pattern

  CREATIVE-WORD was VERB-PAST-PARTICIPLE by ANSWER

- Thus for the question, “Who wrote the Magic School Bus?”, our system would respond with, “The Magic School Bus was written by Joanna Cole”
Knowledge Base Question Answering

- Originally designed to answer questions about information commonly stored in databases, such as baseball statistics
- Modern KBQA systems bring together several techniques we have already looked at in this class
  - Semantic Parsing
    - This is used to take the question from a textual representation to one that a database query could be generated from
  - Relation Extraction
    - To keep knowledge bases up to date without manual population, relation extraction is often used.
Rule Based Semantic Parsing for Question Answering

- Rule based methods are primarily used for questions and relations you expect to encounter a lot.
  - What year was PERSON born?, When was PERSON born?
    - born(PERSON, x)
    - SELECT year FROM born where person = PERSON
  - Who plays James Bond?
    - role(x, James Bond)
    - SELECT actor FROM role where character = “James Bond”

- Are usually manually constructed, so they are expensive, but are usually very accurate.
Semantic Parsing for Question Answering

- Supervised Learning
- Because we are only training over questions, the systems tend to be rather accurate
  - They are often constructed in a bootstrapped fashion
- A good training set would be a set of questions in lambda calculus or SQL, and the corresponding question
  - Dependency parsing and the relations that are produced is a good set of features
  - Additional features might be semantic roles or the primary verb of the sentence.
Evaluation

- For many years, the National Institute of Standards and Technology (NIST, located in Gaithersburg, MD) ran the primer shared task for question answering.
- The Text Retrieval and Evaluation Conference (TREC) had a QA task whose framework is still largely followed to this day.
- The evaluation metric is **Mean Reciprocal Rank (MRR)**.
- Because QA systems often return a list of answers, ranked in some way, we evaluate them based on how far down the list the correct answer is.

\[
MRR = \frac{\sum_{i=1}^{N} \frac{1}{rank_i}}{N}
\]
For the given questions and answer lists, where the correct answer is highlighted, what is the MRR?

Who wrote The Magic School Bus?
**Joanna Cole**, Bruce Degen, Little Richard

What is the capital of Maryland?
Baltimore, **Annapolis**, Rockville

Where is the next Olympics?
Japan, Tokyo, **PyeongChang**
Future Directions

● A majority of QA systems produced thus far have been geared towards factoid-based QA, where the answer is a single piece of information.

● Non-factoid based QA often involves answers that require more text than just a single sentence or phrase.

● How Questions are a good example:
  ○ How to do something?
  ○ How does something work?

● In addition to finding the answers, the different steps or information may be spread out across documents, so synthesis is needed, like in summarization.