Natural Language Processing

Lecture 3: About the Project

Build a Question/Answer System

- Given a Wikipedia article
  - Generate N "good" questions
- Given a Wikipedia article
  - Answer N questions generated from that article

What is a good question?

Pittsburgh (/ˈpɪtsbərɡ/ PITS-burg) is a city in the Commonwealth of Pennsylvania in the United States, and is the county seat of Allegheny County. The Combined Statistical Area (CSA) population of 2,659,937 is the largest in both the Ohio Valley and Appalachia, the second-largest in Pennsylvania after Philadelphia and the 20th-largest in the U.S. Located at the confluence of the Allegheny and Monongahela rivers, which form the Ohio River, Pittsburgh is known as both "the Steel City" for its more than 300 steel-related businesses, and as the "City of Bridges" for its 446 bridges. The city features 30 skyscrapers, two inclines, a pre-revolutionary fortification and the Point State Park at the confluence of the rivers. The city developed as a vital link of the Atlantic coast and Midwest. The mineral-rich Allegheny Mountains made the area coveted by the French and British empires, Virginia, Whiskey Rebels, and Civil War raiders.

Good Questions

- Where is Pittsburgh?
- What is the population of Pittsburgh?
- What is Pittsburgh's nickname?
- How many steel-related businesses are there in Pittsburgh?
- What does the city feature?

What is a bad question?

Pittsburgh (/ˈpɪtsbərg/ PITS-burg) is a city in the Commonwealth of Pennsylvania in the United States, and is the county seat of Allegheny County. The Combined Statistical Area (CSA) population of 2,659,937 is the largest in both the Ohio Valley and Appalachia, the second-largest in Pennsylvania after Philadelphia and the 20th-largest in the U.S. Located at the confluence of the Allegheny and Monongahela rivers, which form the Ohio River, Pittsburgh is known as both "the Steel City" for its more than 300 steel-related businesses, and as the "City of Bridges" for its 446 bridges. The city features 30 skyscrapers, two inclines, a pre-revolutionary fortification and the Point State Park at the confluence of the rivers. The city developed as a vital link of the Atlantic coast and Midwest. The mineral-rich Allegheny Mountains made the area coveted by the French and British empires, Virginia, Whiskey Rebels, and Civil War raiders.

Bad Questions

- What is the capital of France
- What is the 57th letter of the article
- What are Pittsburgh's Three Rivers?

How to find questions

- X is Y → What is Y? (what/why/who/when)
- The X verbs Y → What does the X verb?
- Number-based questions (for article type)

How to answer questions

- What is Y? Look for X is Y
- "extractive" answers
  - Subset of the text
  - Maybe with small changes

Evaluation of questions/answers

- Humans look at them
  - Fluency, reasonableness
- Automatic techniques
  - Will need to rank answer quality
  - Will need to rank fluency of text

Question Asking

- **Input**: text of a Wikipedia article, and an integer n.

- **Output**: n distinct questions about the article. They should be:

  – fluent

  – reasonable

```
./ask article.txt nquestions
```

Question Asking: Development

- We'll give you:
  - Wikipedia articles in five domains
  - Sample questions (generated by your team)

Question Asking: Evaluation

- We choose n (you don't know in advance).
- We'll use your program to generate questions on
  - Some of the articles you had access to
  - Similar articles (same kinds of topics)
  - A completely different type of topic (still Wikipedia)
- Each question will be evaluated:
  - How fluent?
  - How difficult?

Question Answering

- **Input**: text of a Wikipedia article, and a list of questions about the article.

- **Output**: the answers to the questions. The answers should be:

  – fluent

  – correct

  – intelligent-human-like

```
./answer article.txt questions.txt
```

Question Answering: Development

- We'll give you:
  - Wikipedia articles in five domains
  - Sample questions (generated by your team)
  - Sample answers (generated by your team)

Question Answering: Evaluation

- We'll feed your system questions about:
  - Some of the articles you already had access to
  - Similar articles (same kinds of topics)
  - A completely different type of topic (still Wikipedia)
- Each answer will be evaluated:
  - How fluent?
  - How correct?

Initial Tasks

- Form teams
- Build a question generator

Question Generation

- Build a pipeline
  - Analyze an article
  - Segment it in to "sentences"
  - Tokenize each sentence into words
  - Run a part-of-speech tagger on it
  - Find all occurrences of "is" (or some simple verb)
  - Replace subject with appropriate wh-word

Question Generation

- Which wh-word
  - Can you find rules for what/when/who/..
  - Can you exclude the errors

Question Generation

- How good is it?
  - How many does it get right
  - How many are wrong
  - Can you fix these
- Build your own evaluation function

Some Guidelines

- No neural end-to-end systems
  - Your job is to show off what NLP you **explicitly** know
- No external QA systems
  - You can't send it to Google for the answer
- You **can** use existing toolkits
  - NLTK/Stanford Core NLP etc
- You can use NN/word embeddings
  - If they are pretrained and/or fast to train

Administrivia

- Website/Piazza

- Waitlist

- Form teams
  - Choose a team name
  - 4 members in each team
  - Arrange to meet/communicate/pass code